HiLCoE Journal of Computer Science and Technology

December 2022 Volume 09

HiLCoE

School of Computer Science and Technology

HiLCoE Journal of Computer Science and Technology is published once a year by HiLCoE. Individual articles can be accessed and downloaded at https://www.hilcoe.net/journal.

Address all correspondences to: *HiLCoE Journal of Computer Science and Technology*, P. O. Box 25304/1000, Addis Ababa, Ethiopia; e-mail: hjcst@hilcoe.net; Tel. +251 111 56 49 00/ +251 118 68 12 70

Papers for publication are welcome from researchers in Computer Science and Technology. Papers will be peer-reviewed by at least two reviewers who are active researchers in the respective area of each topic of the paper.

Primary Objectives:

- To serve as a forum for the advancement of computer science and technology.
- To contribute to a rapid information and knowledge exchange between researchers in computer science and technology.
- To serve as a forum for postgraduate students in computer science and technology to publish their thesis/research project results.
- To share best practices and case studies from practitioners in the industry that can be sources of new research ideas.

Contents

A Technique to Detect and Categorize COVID-19 Disinformation from Amharic and English Twitter feeds among Ethiopian Twitter Users

Yalew Adimassu,
HilLCoE of School of Computer Science & Technology, yalewadmasu@gmail.com

Eyob Nigussie,
HilLCoE of School of Computer Science & Technology, eyobniguse@gmail.com

Abstract

Since its announcement as pandemic by WHO, COVID-19 virus has been a devastating experience that the world has faced. Many have lost their precious lives to COVID-19 virus. As the virus is not well-known to medical communities during its eruption, many have put forward speculations as to how the virus comes into existence, how it is transmitted, how it could be prevented. Unsubstantiated rumors, fake remedies and falsehoods quickly surfaced social media platforms. Hence; identifying disinformation regarding COVID-19 for a regular person is challenging specially when much is not known and authorities does not have utilities to guide public opinions regarding the virus. If unverified rumors and speculations people hold regarding COVID-19 are shared on social media like twitter, people might subscribe to the ideas embedded in the rumors and might apply to the real world by considering these rumors a legitimate claim. This thesis work has proposed a technique that will detect disinformation twitter feeds/claims. The detection model works for both Amharic and English languages. Amharic dataset contains 205 non-disinformation claims and 196 disinformation claims; similarly, English dataset included 568 non-disinformation and 585 disinformation claims. The developed technique has been deployed on Flask web application to serve as COVID-19 disinformation detection platform. The developed detection technique has F-1 score of 90.48% for English dataset while the same model recorded F-1 score of 91.4% for Amharic dataset. This thesis work could be expanded further by incorporating other social media sources like, Facebook, WhatsApp, Telegram, etc. as the source COVID-19 conversations to further enrich the datasets.

Keywords: COVID-19 ·Coronavirus ·Disinformation Detection · Logistic regression · Amharic tweets · Tweet classification

1. Introduction

COVID-19 pandemic is one of the most devastating experiences of human history which has brought significant and multifaceted social, economic and political impacts all over the world. As it has been explained in [1], COVID-19 pandemic is an unprecedented crisis unlike any since the end of the second world war. The virus was first detected in Wuhan, China, in late 2019, then

rapidly transmitted to all Chinese provinces and many other countries by late January 2020[2]. WHO has declared the COVID-19 outbreak as a pandemic on March 11, 2020. According to WHO, as of 18th of April 2021, there are over 140 million confirmed cases of COVID-19 and over 3 million people have lost their lives to the virus [3]. Out of the reported number by WHO, Africa accounts for about 3.2 million confirmed COVID-19 cases. As of April 2, 2021, Ethiopian Public Health Institute (EPHI) public reports shows there are more than 200,000 confirmed COVID-19 cases in Ethiopia and around 2,800 people have lost their lives due to COVID-19 pandemic[4]. As [5] put it, disinformation in general has its footing since 1835. According to [6], disinformation is defined as claim of fact that is currently false due to lack of scientific evidence that supports the stated claim. Twitter is a social media where people can get instant updates and messages regarding contemporary subjects of interest. As per the authors in [6], 25% of sampled contents on twitter were found to be untrue or misleading. As it has been cited in [7], WHO Director-General Dr. Tedros Adhanom said "We're not just fighting an epidemic; we're fighting an infodemic", at the Munich Security Conference on Feb 15,2020 , stressing the impact of disinformation being disseminated regarding COVID-19. Several efforts are being made by researchers to develop a technique to detect disinformation from COVID-19 related claims. A common theme in developing COVID-19 disinformation detection is by utilizing a validated dataset. The authors in

[8] has provided a novel way of detecting COVID-19 related fake news by integrating machine learning techniques with statistical features. In addition to the exact tweets or news under investigation, [8] has considered username handle, URL domains, author of the news and news source to create the proposed framework to detect and identify fake news surrounding COVID-19 discussions in twitter. There are various papers done to detect disinformation in languages other than English. For example; [9]has developed a technique to detect disinformation from Arabic tweets. [9] have proposed a two-step process named claim-level verification and tweet level verification in detecting COVID-19 disinformation. According the work in [9], claim-level verification is defined as given short free-text claim in 1 or 2 sentences and its corresponding relevant tweets, prediction is made whether the claim is true or false whereas tweet-level verification is defined as given a tweet containing a claim, detect whether it is true or false. This thesis work attempts devise disinformation detection technique for Amharic and English language on the basis of keyword repetition within the given claim.

2. Related Works

As COVID-19 pandemic is global challenge that affected all continents, several researches are being undertaken in different contexts regarding COVID-19 disinformation. The authors in [10] developed a methodology to extract top negative and positive words used in COVID-19 related tweets spanning from 25th of march to 14th of April 2020 for twitter users in India. The developed methodology has enabled to capture psychological state of the twitter users in India. The authors in [11] have conducted comparative analysis of machine learning algorithms in identifying sentiment polarity on twitter. The authors in [12] performed sentiment analysis on the impact of COVID-19 in social life using BERT model. In developing the model [12] has used twitter datasets which includes tweets from people living in India and around the world. The authors in [2] have developed a methodology that enabled forecasting COVID-19 activities in Chinese province by utilizing internet searches, news alerts, and estimates from mechanistic models. According to the

experimentations done in [2], the developed methodology's predictive power outperforms a collection of baseline models. The authors in [13] performed a comparative research on text preprocessing methods for tweet sentiment analysis and the findings showed that the performance of sentiment classifiers improved by expanding acronyms in the dataset and by including expanded negation words instead of contracted form of negation words. The authors in [14] studied how machine learning techniques can be implemented in analyzing pattern of COVID-19 transmission in India with the help of authoritative datasets which has led to successful analysis of COVID-19 transmission trend in the world that might help decision makers to better position themselves to combat the virus. The authors in [15] has developed a framework to analyze dynamics of socialmedia opinions regarding the first introduction of vaccination in November 2020 in UK. [15] has compared classical machine learning models against the given vaccination opinion dataset. [16] proposed a new contact tracing methodology which utilizes machine learning algorithms and smartphones activity logs to trace COVID-19 infected patients. [17] has developed a technique to remove irrelevant COVID-19 tweets from a given COVID-19 tweet dataset which enabled to identify widely discussed topics and underlying discussion themes in a given dataset. The methodology developed by the authors in [17] further enabled to detect temporal changes in public discourse as key events unfold during the course of the pandemic. This thesis shall base its footing on the existing related works to develop a technique that allows to detect COVID-19 disinformation from Amharic and English language tweets.

3. Solution Design

Disinformation detection techniques in the existing literatures has been investigated to see the trend in disinformation detection in general and COVID-19 disinformation detection in particular to devise technique for detection of COVID-19 disinformation from Amharic and English tweets.

3.1. Dataset Summary of Amharic and English dataset

Verified COVID-19 related claims from authoritative sources has been utilized to compile datasets that are classified as either be disinformation or non-disinformation depending on validity of the message being conveyed in the tweets. The following table summarizes English dataset sources along with number of disinformation and non-disinformation claims.

Table 1: English dataset summary

Sources	No. of Disinformation claims	No. of non-disinformation claims
CoVID19-FNIR[18]	456	514
WHO[19]	0	54
AfriCheck[20]	42	0
Politifact[21]	87	0
Total	**585**	**568**

The following table summarizes Amharic dataset sources along with number of disinformation and non-disinformation claims.

Table 2: Amharic dataset summary

Source	No. of Disinformation claims	No. of Non-disinformation claims
EPHI – Guidelines	0	66
EPHI- Twitter account	0	29
MoH Twitter account	0	56
WHO[19]	0	54
AfricaCheck[20]	42	0
Poulitifact[21]	87	0
JNS[22]	67	0

3.2. Proposed Solution

3.2.1. System Input-output Scheme

COVID-19 disinformation detection model takes training dataset to learn about the tweet features within the given texts in the training dataset. Once the model is trained using the training dataset it can detect whether a given claim in the test set is disinformation ("Label-1") or non-disinformation("Label-0"). Figure 1. shows a simplified depiction of data input-output for COVID-19 disinformation detection model.

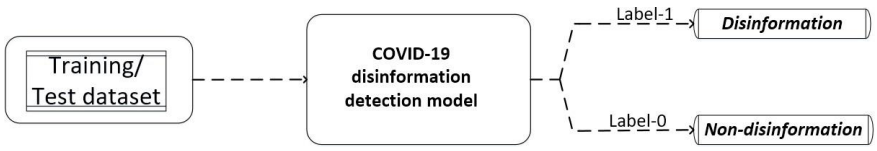


Figure 1. Simplified Data input-output

As it can be inferred from Figure 1., COVID-19 disinformation detection platform takes training dataset as in input to learn the features within the dataset. Once training phase is completed the model shall be tested using the testing dataset to judge whether the given claim in the testing dataset is disinformation or non-disinformation. From the functionality perspective the entire COVID-19 disinformation detection technique enables to determine whether the given COVID-19 related tweet is disinformation or not. As it is briefly depicted in Figure 2, COVID-19 disinformation detection model determines whether the given claim is disinformation or not and notify the requester accordingly by stating whether the given COVID-19 related claim is "disinformation" or "non-disinformation" along with the accuracy of the detection.

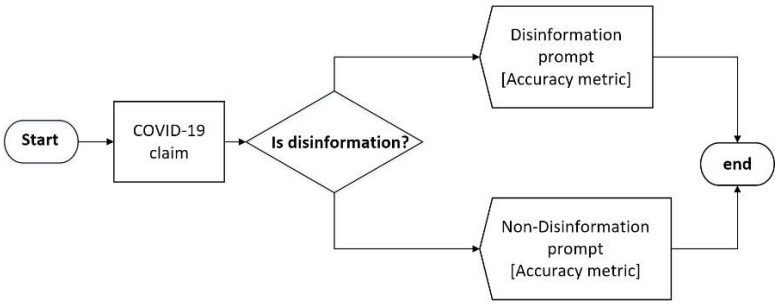


Figure 2. COVID-19 disinformation detection decision tree

3.2.2. End to end disinformation detection workflow

Various steps are involved in determining the truthfulness of a given tweet claim. The process and steps involved in determining disinformation for Amharic and English dataset is slightly different. This is due to the fact that English and Amharic written texts are structurally and semantically different. Therefore, the workflow involved processing Amharic and English texts are separately handled. Figure 3 and 4 depicts disinformation detection steps for English and Amharic language tweets respectively.

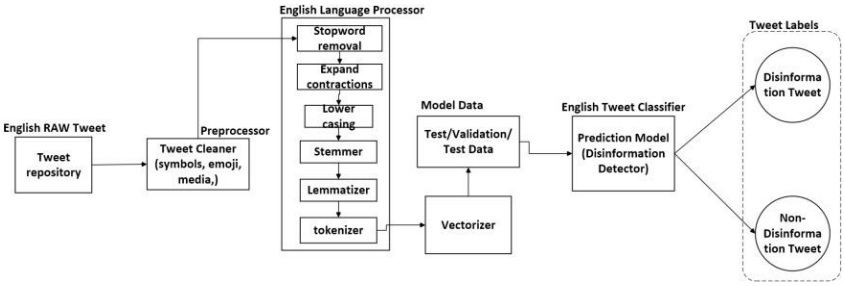


Figure 3. End-to-end disinformation detection work for English tweets

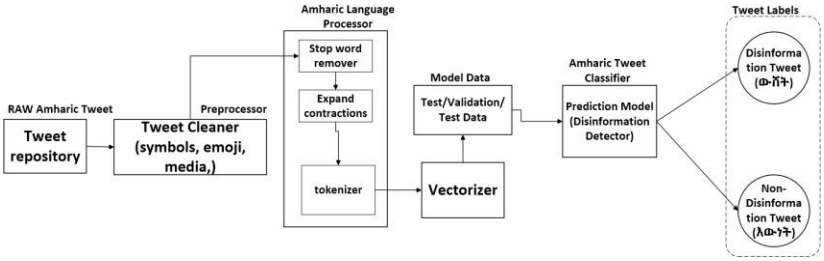


Figure 4. End-to-end disinformation detection work for Amharic tweets

Following is a brief description of disinformation detection steps shown in figure 3 and 4. Stemmer, lemmatizer and lowercasing steps are not applicable to Amharic dataset.

- **Preprocessing**: Removes unwanted contents within the tweet. Unnecessary contents include emojis, symbols and media contents
- **Expand contraction**: Contracted words are changed into the expanded form.
- **Stop word removal**: Commonly used stop words English language like, as, is, are etc., has been be removed without altering the underlying message being conveyed by the tweet.
- **Lowercasing**: for the sake of uniformity all capital letters within the tweet has been converted to lower case letters.
- **Stemmer**: Root words of derived words has been used
- **Lemmatizer**: Dictionary equivalent of a given word has been obtained
- **Tokenizer**: Words in their original form has been tokenized for further processing.
- **Vectorizer**: Words has been represented using numerical vector for further processing.

- **Dataset splitting**: The dataset has been divided into 80/20 ratio for training and test set
- **Model Building:** The best performing model has been chosen.

3.3. Experimentation and Evaluation

As indicated by [23], selection of the best model relies on the model's performance on the testing set. There are various widely adopted theories as to how to proportionate percentage of data should be split among training, validation and testing. For example, authors in [24] has allocated 60/20/20 for training, validation and testing. As it has been discussed in [25] having validation set allows a room for model adjustment by adding or removing features. As it has been indicated in [24] it is customary to reserve 80% of the dataset for training where as 20% of the dataset shall be used for training and testing. The authors in [26] further indicated training/test datasets can be divided as 60/40, 70/30, or 80/20, even 90/10 separation is possible if the dataset is relatively large. In this thesis 80/20 split scheme has been selected to train and test the developed model.

3.3.1. Experimentation on Amharic dataset

Following 80/20 training and testing segregation rule proposed by [24] and [26], quantities of training and testing dataset distribution is shown in Table 3.

Table 3: Amharic dataset model training and testing split

Total entries in the dataset	**401**	
Training set(80%)	No. of X_train entries	320
	No. of y_train entries	320
Test Set(20%)	No. of X_test entries	81
	No. of y_test entries	81

According Table 3, out of 401 Amharic dataset entries 80% of the entries has been utilized to train the COVID-19 disinformation detection model while the rest of the entries (i.e. 20%) has been used to test the efficacy of the trained COVID-19 disinformation detection model.

Table 4, contains various recorded performance evaluation metrics of COVID-19 disinformation detection technique for Amharic dataset on the basis of SVM, logistic regression, gradient boost and random forest machine learning algorithms.

Table 4: Disinformation detection technique evaluation metrics for Amharic dataset

Model	**Metrics**	**Values**	**Values in %**
SVM	Accuracy on test set	0.716	71.60%
	F-1 score	0.6844	68.44%
Logistic Regression	Accuracy on test set	0.9136	91.36%
	F-1 score	0.9137	91.37%
Gradient Boost	Accuracy on test set	0.7654	76.54%
	F-1 score	0.7661	76.61%
Random Forest	Accuracy on test set	0.7901	79.01%
	F-1 score	0.7895	78.95%

Running COVID-19 disinformation detection model on Amharic testing dataset have shown different results for each underlying machine learning models. As per the records in Table 4,

COVID-19 disinformation detection technique on the basis of logistic regression has recorded the highest accuracy value which is 91.36%. In a similar comparison accuracy of the developed technique on the testing dataset on the basis of SVM has relatively poor value which is 71.6% compared to the disinformation detection technique on the basis gradient boost and random forest. The accuracy of disinformation detection technique on the basis of gradient boost and random forest is 76.54% and 79.01% respectively. Comparing F-1 score values of COVID-19 disinformation detection model shown in Table 4 shows that disinformation detection model on the basis of logistic regression has the maximum F-1 score value which is 91.37% while disinformation detection model on the basis of SVM has the lowest F-1 score value which is 68.44%. In a similar comparison, F-1 score of disinformation detection model on the basis of random forest and gradient boost nearly equal (i.e., 76.61% for gradient boost, 78.95% for random forest). In summary COVID-19 disinformation detection model on the basis of logistic regression has relatively maximum F-1 score (i.e. 91.37%) and accuracy on the test set (i.e., 91.36%) compared to disinformation detection technique on the basis of SVM, gradient boost and random forest.

3.3.2. Experimentation on English dataset

Following 80/20 training and testing segregation rule proposed by [24] and [26], quantities of training and test dataset distribution is shown in Table 5.

Table 5: English dataset model training and testing split

Total dataset	**1153**	
Training set(80%)	No. of X_train entries	922
	No. of y_train entries	922
Test Set(20%)	No. of X_test entries	231
	No. of y_test entries	231

According to Table 5, out of 1,153 English dataset entries 80% of the entries has been utilized to train the COVID-19 disinformation detection model while the rest of the entries (i.e. 20%) has been used to test the efficacy of the trained COVID-19 disinformation detection model.

Table 6., contains various recorded evaluation metrics of COVID-19 disinformation detection technique for English dataset on the basis of SVM, logistic regression, gradient boost and random forest machine learning algorithms.

Table 6: Disinformation detection technique evaluation metrics for English dataset

Model	**Metrics**	**Values**	**Values in %**
SVM	Accuracy on test set	0.883116883	88.31%
	F-1 score	0.883195952	88.32%
Logistic Regression	Accuracy on test set	0.904761905	90.48%
	F-1 score	90%	90.48%
Gradient Boost	Accuracy on test set	0.883116883	88.31%
	F-1 score	0.883173838	88.32%

Random Forest	Accuracy on test set	0.891774892	89.18%
	F-1 score	0.891722155	89.17%

Running COVID-19 disinformation detection model on English testing dataset have shown different results for each underlying machine learning models. As per the records in Table 6, COVID-19 disinformation detection technique on the basis of logistic regression has recorded the highest value of accuracy and F-1 score which is 90.48%. In a similar comparison from Table 6, accuracy and F-1 score of the developed technique on the testing dataset on the basis of random forest is 89.18%, 89.17% respectively. The accuracy of detection model on the basis of SVM and gradient boost is similar which is 88.31%. Similarly, the F-1 score of detection model on the basis of SVM and gradient boost is equal which is 88.32%. Evaluation results for both Amharic and English datasets is above 90%. The developed disinformation detection technique on the basis of logistic regression has learned widely used keywords in each disinformation and non-disinformation datasets. Frequently appearing keywords within a given claim will be checked against the trained model to determine whether the given claim is disinformation or not.

4. Discussion

This thesis has explored existing literatures that are available surrounding disinformation detection. Twitter has been used as a target social platform from which tweets are collected and analyzed. Amharic and English language support in detecting COVID-19 disinformation has been central work of this proposed technique. The proposed technique has been implemented to allow users to access the platform to check the validity a given Amharic or English language tweets. To complement an effort to fight COVID-19 disinformation, the developed COVID-19 disinformation detection technique has been operationalized as web application that can be used as a web application to check the validity of COVID-19 rumors circulating on twitter. Any allegedly suspicious tweets can be checked by entering the tweet onto the web application and the disinformation detection platform checks truthfulness of the tweet according to the currently available authoritative health information sources. Information regarding the virus varies overthe period of time as new COVID-19 virus variant appears and therefore new claims emerge which requires investigating the new claims' validity. The developed technique can adapt to new emerging COVID-19 claims by re-training the model using newly available information regarding COVID-19 pandemic. This adaptability enables decision makers to provide up to date and factual COVID-19 related information to a general public.

5. Conclusion

Combating COVID-19 pandemic is highly dependent on creating awareness regarding how the virus is transmitted, what the safety guidelines are and what are the current authorized health remedies available. This requires fighting disinformation efforts on social media where most people get instant messages which may expose them to unsubstantiated claims, fake remedies and falsehoods. This thesis work has attempted to devise a technique to detect COVID-19 disinformation from Amharic and English tweets. In order to develop the model online and offline tweet sources have been utilized for both Amharic and English languages. The evaluation result has shown that, COVID-19 disinformation detection technique on the basis of logistic regression technique performed best in detecting disinformation for both Amharic and English language tweet.

The developed technique has been trained using Amharic and English datasets which has been collected from online and offline authoritative sources. All existing works so far did not include dual language support in detecting COVID-19 disinformation. As of conducting this thesis work COVID-19 disinformation detection from Amharic texts has not been done so far. In summary this thesis work has attempted to make the following contributions:

- Developed a technique that allowed detection of COVID-19 disinformation from Amharic tweets by utilizing Logistic regression as an underlying technique
- COVID-19 safety guidelines have been directly scrapped and a dataset has been constructed from relevant texts from the guideline documents. This has contributed to authenticity of disinformation detection effort
- Disinformation on the basis of keywords learning from the dataset has been used as scheme to detect disinformation
- Dual language support has been realized by combining COVID-19 disinformation systematically integrating disinformation detection techniques for English and Amharic languages.
- COVID-19 disinformation detection technique has been deployed as web application to show case the operational state of the developed COVID-19 disinformation detection technique.

6. Future Work

The aim of this thesis work has been to develop a technique that enables detection of COVID-19 disinformation from Amharic and English tweets. Even if an effort has been made to make this thesis work as extensive as possible, there are still areas where improvements can be made by researchers working on COVID-19 disinformation detection. This thesis work has only considered tweet texts as a feature in constructing the datasets for both Amharic and English languages. However other features within a given tweet like, photo, video, can also be considered in identifying disinformation. This thesis work could be expanded further to include more COVID- 19 related claims from other social media sources like Facebook, Instagram, Telegram, WhatsApp, etc. to improve the quality and efficacy of the COVID-19 disinformation detection technique. Such diverse inclusion of COVID-19 related claims could lead to covering wide range of opinions/claims regarding COVID-19 which can be used to train COVID-19 disinformation detection model. Additional machine learning techniques apart from logistic regression, SVM, gradient boost and random forest could be explored to further improve the performance of COVID- 19 disinformation detection. A support for additional language is a possibility if one wants to make this technique to be utilized widely. The developed COVID-19 disinformation detection technique can be operationalized in such a way that can complement health professionals and policy makers' decision making process by providing real time COVID-19 disinformation conversation trends while combating COVID-19 disinformation on twitter.

References

[1] United Nations, "SOCIO - ECONOMIC IMPACT of COVID-19 IN ETHIOPIA," no. May, 2020.

[2] D. Liu *et al.*, "A machine learning methodology for real-time forecasting of the 2019-2020 COVID-19 outbreak using Internet searches, news alerts, and estimates from mechanistic models.," *ArXiv*, no. d, 2020.

[3] World Health Organization, "COVID-19 Weekly Epidemiological Update 22," *World Heal. Organ.*, no. December, pp. 1–3, 2021.

[4] EPHI and MoH, "Public Health Emergency Operation Center (PHEOC), Ethiopia COVID-19 Pandemic Preparedness and Response in Ethiopia," *EPHI Wkly. Bull.*, vol. 39, pp. 1–21, 2021.

[5] H. Kudarvalli and J. Fiaidhi, "Detecting Fake News Using Machine Learning Algorithms," *2021 Int. Conf. Comput. Commun. Informatics, ICCCI 2021*, 2021, doi: 10.1109/ICCCI50826.2021.9402470.

[6] R. Kouzy *et al.*, "Coronavirus Goes Viral: Quantifying the COVID-19 Misinformation Epidemic on Twitter," *Cureus*, vol. 12, no. 3, 2020, doi: 10.7759/cureus.7255.

[7] J. Zarocostas, "How to fight an infodemic," *Lancet (London, England)*, vol. 395, no. 10225, p. 676, 2020, doi: 10.1016/S0140-6736(20)30461-X.

[8] H. Ito and B. Chakraborty, "Social media mining with dynamic clustering: A case study by COVID-19 Tweets," *2020 11th Int. Conf. Aware. Sci. Technol. iCAST 2020*, 2020, doi: 10.1109/iCAST51195.2020.9319496.

[9] J. Nguyen and R. Chaturvedi, "Quarantine Quibbles: A Sentiment Analysis of COVID-19 Tweets," *11th Annu. IEEE Inf. Technol. Electron. Mob. Commun. Conf. IEMCON 2020*, pp. 346–350, 2020, doi: 10.1109/IEMCON51383.2020.9284875.

[10] R. Kaur and S. Ranjan, "Sentiment Analysis of 21 daysCOVID-19 Indian lockdown tweets,"

Int. J. Adv. Res. Sci. Eng., vol. 9, no. 7, pp. 37–44, 2020.

[11] M. I. Sajib, S. Mahmud Shargo, and M. A. Hossain, "Comparison of the efficiency of machine learning algorithms on twitter sentiment analysis of pathao," *2019 22nd Int. Conf. Comput. Inf. Technol. ICCIT 2019*, pp. 1–6, 2019, doi: 10.1109/ICCIT48885.2019.9038208.

[12] M. Singh, A. K. Jakhar, and S. Pandey, "Sentiment analysis on the impact of coronavirus in social life using the BERT model," *Soc. Netw. Anal. Min.*, vol. 11, no. 1, pp. 1–11, 2021, doi: 10.1007/s13278-021-00737-z.

[13] Z. Jianqiang and G. Xiaolin, "Comparison research on text pre-processing methods on twitter sentiment analysis," *IEEE Access*, vol. 5, pp. 2870–2879, 2017, doi: 10.1109/ACCESS.2017.2672677.

[14] P. Mars, J. R. Chen, and R. Nambiar, "Learning Algorithms," *Learn. Algorithms*, no. Icosec, pp. 65–71, 2018, doi: 10.1201/9781351073974.

[15] L. A. Cotfas, C. Delcea, I. Roxin, C. Ioanăş, D. S. Gherai, and F. Tajariol, "The Longest Month: Analyzing COVID-19 Vaccination Opinions Dynamics from Tweets in the Month following the First Vaccine Announcement," *IEEE Access*, vol. 9, pp. 33203–33223, 2021,

doi: 10.1109/ACCESS.2021.3059821.

[16] A. Narzullaev, Z. Muminov, and M. Narzullaev, "Contact Tracing of Infectious Diseases Using Wi-Fi Signals and Machine Learning Classification," *IEEE Int. Conf. Artif. Intell. Eng. Technol. IICAIET 2020*, 2020, doi: 10.1109/IICAIET49801.2020.9257812.

[17] A. Agarwal, P. Salehundam, S. Padhee, W. L. Romine, and T. Banerjee, "Leveraging Natural Language Processing to Mine Issues on Twitter during the COVID-19 Pandemic," *Proc. - 2020 IEEE Int. Conf. Big Data, Big Data 2020*, no. November, pp. 886–891, 2020, doi: 10.1109/BigData50022.2020.9378028.

[18] J. A. Saenz, S. R. K. Gopal, and D. Shukla, "Covid-19 Fake News Infodemic Research Dataset (CoVID19-FNIR Dataset)," *IEEE Dataport*, 2021.

[19] WHO, "Coronavirus disease (COVID-19) advice for the public: Mythbusters." [Online]. Available: https://www.who.int/emergencies/diseases/novel-coronavirus-2019/advice-for-public/myth-busters. [Accessed: 13-Apr-2021].

[20] AfricaCheck, "LIVE GUIDE: All our coronavirus fact-checks in one place - Africa Check." [Online]. Available: https://africacheck.org/fact-checks/reports/live-guide-all-our-coronavirus-fact-checks-one-place. [Accessed: 13-Oct-2021].

[21] Politifact, "LATEST FALSE FACT-CHECKS IN CORONAVIRUS." [Online]. Available: https://www.politifact.com/factchecks/list/?category=coronavirus&ruling=false. [Accessed: 13-Jan-2022].

[22] N. Bariletto, J. Oledan, and L. Z. V, "COVID Misinformation Codebook V3.0, 2020 This Version Updated: V3.0 Sept 4," vol. 1, pp. 1–11, 2020.

[23] C. Nantasenamat, "How to Build a Machine Learning Model | by Chanin Nantasenamat | Towards Data Science," 2020. [Online]. Available: https://towardsdatascience.com/how-to-build-a-machine-learning-model-439ab8fb3fb1. [Accessed: 10-Feb-2020].

[24] P. Patwa *et al.*, "Fighting an Infodemic: COVID-19 Fake News Dataset," *Commun. Comput. Inf. Sci.*, vol. 1402 CCIS, pp. 21–29, 2021, doi: 10.1007/978-3-030-73696-5_3.

[25] Google LLC, "Machine Learning Crash Course: Validation Set: Another Partition," 2020. [Online]. Available: https://developers.google.com/machine-learning/crash-course/validation/another-partition. [Accessed: 13-Jan-2022].

[26] S. Raschka, "Model Evaluation, Model Selection, and Algorithm Selection in Machine Learning," 2018.

An Algorithm for Detection of Blackhole Attack on RPL

Meron Mekonnen
HiLCoE School of Computer Science and Technology
gatomeron@gmail.com

Taye Abdulkadir
Assistant Professor, HiLCoE School of Computer Science and Technology
tabdulkadir@gmail.com

Abstract

The Internet of things (IoT) is an emerging technology, which seems to have a promising future in linking the real and virtual world and hence changing the way we experience the world. As its implementation increases, attacks performed on it will also increase. Since it is connected with the physical world, the attacks can result in a serious damage. Consequently, it is required to provide protection mechanisms against the attacks.

Blackhole attack is one of the attacks, which is performed on RPL (Routing Protocol for Low- Power and Lossy Networks). In RPL network, the nodes connect to the network forming a tree topology. There are parents and children nodes, and packet is transferred only from children to parent or parent to children until it reaches its destination. In a blackhole attack, attacker node drops packets passing towards it, resulting in the packets from and towards its children not reaching their destination. This attack can result a serious consequence especially where real time data transfer is required.

In this work, an algorithm is designed for detecting the existence of blackhole attack. After an attack is detected, the attacker node (blackhole node) is identified and removed from the network. Blackhole attack is simulated using the Cooja simulator and the provided algorithm is tested by using *packet delivery rate* and *energy consumption* as evaluation metrics. Ensuring the improvement of packet delivery rate is vital because blackhole node drops packets. Additionally, IoT contains resource constrained devices and hence energy consumption has to be as minimal as possible.

The provided algorithm has successfully detected the existence of blackhole attack and the blackhole nodes are successfully identified and removed from the network. The simulation shows that the designed algorithm has improved the packet delivery rate without resulting higher energy consumption for a network of up to 22 nodes.

Keywords – *IoT, RPL, Blackhole attack, Detection*

1. Introduction

In the Internet of things (IoT), every day physical objects are connected to the Internet and are able to send and receive data. These objects generate data and communicate with each other as well as with their physical environment. Sensors play important role in IoT since they are the devices collecting heterogeneous data from the physical environment. They are resource- constrained devices with limited memory, processing capacity and battery power. Consequently, they use routing protocols designed for resource constrained devices.

RPL is a distance vector IPv6 routing protocol for low power and lossy networks [2] [3] [4]. Low power and lossy networks contains nodes which often operate in harsh and remote environment [2]. The term "lossy" refers to the unreliable nature of the communications in these networks [2]. RPL creates a routing topology in the form of a destination-oriented directed acyclic graph (DODAG) [2] [3] [4]. Directed acyclic graph is a graph that is not cyclic which means that it is not possible to start at one node and traverse the entire graph or come back to the same node by traversing the edges. The DODAG is a directed graph without cycles, oriented towards the root node which is the LBR (Low-power Wireless Personal Area Networks (LoWPAN) Border Router) [2]. Traffic flows from each node towards its immediate parent until it reaches the root and vice versa. There are various attacks performed on RPL and blackhole attack is one of these attacks [5].

In a blackhole attack, attacker node drops packets passing through it silently. Nodes oblivious of the malicious node, keep sending their packets while the malicious node keep receiving and dropping the packets like a hole which sucks in everything [5].

In the RPL protocol, since the sending node wouldn't expect confirmation message from the destination node, blackhole attack could remain in the network undetected. However, its impact can result in serious consequences and it needs to be identified and mitigated. This work focuses on providing an algorithm to identify and mitigate blackhole attack on the RPL routing.

2. Related Work

F. Ahmed, et al [6] have proposed a mitigation technique for blackhole attack in LLNs using local decision and global verification process. In the local decision process, nodes overhear data packets transmitted by their neighbors and they observe the communication behavior of their neighbors. If a node notices misbehaving activity in its neighbor, it counts the misbehaving activity and compares it to a given threshold value. If the count of misbehaving activity exceeds the threshold value, the neighbor is judged to be a suspicious node. When the threshold value of the suspected node is exceeded, the neighboring node which is the verification node initiates the global verification process. In the global verification process, the verification node sends a request to the root node to verify the suspicious node. The root node checks if the suspicious node is a blackhole node or not. If the suspicious node is found to be blackhole, the verification node broadcasts the blackhole node ID to its neighbors. The child nodes of the blackhole node will then avoid using it and will use another node to transfer data packet.

In order to identify blackhole nodes, each node in the network must actively listen to packets exchanged between their neighbors and observe the communication behavior of their neighbors. This results in high energy consumption for the resource constrained IoT devices which is a drawback on the provided solution [6].

J. Jiang, et al [7] have proposed a defense mechanism for blackhole attack on RPL. They proposed for the root node to actively track the packets instead of each node monitoring its neighbor's packet which removes the high energy consumption problem due to neighbor packet monitoring in [6]. They used a sequence number in each node's message which

increments as a message is sent and which is stored at the root. A counter variable is used at the root which records the number of packets sent out by a node and successfully received by the root. By using the stored sequence number and counted value, the packet loss rate is determined. Nodes with low packet delivery ratio are determined as blackhole nodes. Once the root node identifies a blackhole node, it broadcasts the message to the network. When a node receives the message, it broadcasts the message to its neighbors and if the blackhole node is its parent, it removes the blackhole node from its parents list. By this process the blackhole node is isolated by its children.

In [7], for determining the packet loss rate of a node, its packet is required to reach the root at some point so that packet loss rate of a node could be determined. This solution can be applicable for a blackhole node that drops some packets. However, if the blackhole node drops all the packets, then a node's packet is not going to reach the root, which means that it won't be possible to determine packet loss rate. Additionally, the node with low packet delivery ratio is identified as malicious node. However, having a low packet delivery ratio does not always indicate a blackhole node. A legitimate node could have a low packet delivery ratio if its children sends packet at a lower rate than the other nodes in the network. In this work, packet loss can be determined without requiring for a packet to reach the root. Furthermore, blackhole node identification is done not only by considering possible packet loss but also by considering that there might exist nodes that sends packets at lower frequency than the other nodes in the network, so that these nodes would not be misclassified as having packet loss.

3. The Proposed Solution

In order to detect the presence of blackhole attack on RPL and reduce its impact, an algorithm is provided in this work. The task of the algorithm is to detect blackhole attack on RPL and remove the attacker node from the network. When sending packets, nodes do not expect to receive confirmation for their packet reaching its destination. As a result, they remain oblivious if a blackhole node is dropping their packets which is why blackhole attack succeeds. In the provided algorithm, a sequence number is used in the sent messages to identify packet loss. The sequence number starts from one and increments as the node sends messages. Since IoT contains resource-constrained devices, it would not be possible to store the sequence number of each node on every other node. As a result, the root node is used for keeping the sequence numbers of all the nodes. The sequence number stored in the root is updated when a new packet is received from a node. Based on the stored sequence number on the root, each node's sequence number is checked relative to the other nodes, and it is determined whether the node is delayed or not.

"Sequence delay percentage (SDP)" is used to check sequence number of a node relative to the other nodes in the network. By comparing the sequence number of a node to the sequence number of all the other nodes in the network, the sequence delay percentage gives from what percent of the nodes in the network, is a node's sequence number low (delayed). Even if there is no current sequence number received from a node, the root can still compare the sequence number of the node relative to the other nodes using the previously stored sequence number of the node. This can be helpful when there is no packet to be received from a node

either due to the presence of blackhole attack or the node being a node which sends packets at lower frequency than the other nodes.

The steps used for computing the sequence delay percentage is discussed below

- *Perform the SD_{iff} of the node to be checked relative to node j , where node j represents all the other sender nodes excluding the node being checked*

$SD_{iff} = S_j - (I * S_c)$

Where SD_{iff} = sequence difference relative to node j S_j = sequence number of node j

S_c = sequence number of the node being checked I = interval

- *Count the number of nodes whose SD_{iff} is greater than zero*

- *Determine the sequence delay percentage for the node*

$$ctSDP = \frac{}{n-1} x100$$

Where SDP = sequence delay percentage

ct = count of number of nodes whose SD_{iff} is greater than zero n = total number of nodes excluding the root node The sequence difference of a node is its sequence number difference relative to the other node it is being checked to. It is used to determine whether a node has a smaller sequence number or not relative to the other node which is used for comparison. Positive SD_{iff} value indicates that the node being checked has a lower sequence number below the acceptable limit and hence the node is possibly delayed relative to the node it is being checked to. On the other hand, negative SD_{iff} indicates that the node being checked is not delayed relative to the node used for checking.

The count (ct) value determines the number of positive sequence difference values obtained. Higher count value for a node means a node's sequence number is less than its expected sequence value relative to large number of nodes.

3.1 Proposed algorithm in this work

Module1: *Sequence check of nodes*

```
Let nd = number of nodes excluding the root
for i=0 to nd-1 do
    ct=0
    for j=0 to nd-1 do
        SDiff = Sequence[j]-Interval*Sequence[i]
        if SDiff > 0
            ct=ct+1
        end if
    end for
    SDP[i] = (ct/(n-1))*100
```

```
11  end for
```

Module2: *Identify delayed nodes*

```
1  for i=0 to nd-1 do
2        if SDP[i]>50
3             delayed_node[i]=node_id(i)
4             delayed_time[i]=delayed_time[i]+1
5        else
6             delayed_node[i]=0
7             delayed_time[i]=0
8        end if
9  end for
```

Module3: *Send message to delayed nodes*

```
1  for i=0 to nd-1 do
2       if delayed_node[i]!=0
3            if delayed_time[i]>0 and delayed_time[i]<=3
4                 client_ipaddr=ip_return(delayed_node[i])
5                 uip_udp_packet_sendto((server_conn, msg,
                  sizeof(msg),client_ipaddr,
                  UIP_HTONS(UDP_CLIENT_PORT)) //This is is part of
                  Contiki OS
6            else if delayed_time[i]>3
7                 attacked_nodes_ct = attacked_nodes_ct + 1
8            end if
9       end if
10 end for
11 if attacked_nodes_ct!=0
12      blackhole_node_children() //Module4
13      blackhole_node_id() //Module5
14 end if
```

Module4: *Identify attacked nodes*

```
1  attacked_ct=0
2  for i=0 to nd-1 do
3       if delayed_node[i]!=0 and delayed_time>3
4            attacked_node[attacked_ct]=delayed_node[i]
5            attacked_ct=attacked_ct+1
6       end if
7  end for
```

Module5: *Identify blackhole node*

```
1  attacked_ct=get_attacked_ct()
2  for i=0 to attacked_ct-1 do
3        ancestor_get(attacked_node[i]) //gives ancestors of a node starting
       from the node's parent
4        ancestor=attacked_node[i]
5        for j=0 and ancestor!=1
6             ancestor=ancestor_table[i][j]
7             if SDP[ancestor]<50
8                  blackhole_node[blackhole_ct]=ancestor
9                  blackhole_ct = blackhole_ct + 1
10                 ancestor=1
11            end if
12       end for
13 end for
```

Module6: *Remove blackhole node*

```
1  blackhole_ct=get_blackhole_ct()
2  for i=0 to blackhole_ct-1 do
3        client_ipaddr=ip_return(bh_node_neighbour[blackhole_node[i]])
4        uip_udp_packet_sendto(server_conn, msg,
         sizeof(msg),client_ipaddr,    UIP_HTONS(UDP_CLIENT_PORT))
         //This is part of Contiki OS
```

3.2 Blackhole attack detection

The sequence delay percentage is determined for every node so that to identify whether there is packet loss or not. Higher SDP means the checked node's sequence delay percentage is higher compared to the sequence number of the other nodes in the network, which might indicate packet loss. However, having higher SDP does not always mean there is a packet loss. It might be because of the node being checked sending packets at lower frequency than the other nodes. Furthermore, packet loss necessary might not mean the presence of blackhole attack. The packet loss might be due to the lossy nature of the network. Which means that a packet might be lost since the network is unreliable. Consequently, 50% is used as a threshold value, and hence, for a node to be considered delayed, it needs to have a sequence delay percentage of greater than 50% which means that its sequence is delayed compared to more than half of the other nodes.

After a node is identified as delayed, a message is sent by the root to the delayed node to confirm whether the higher SDP value is due to the presence of blackhole attack or due to the required task of the node. The root then expects to receive a reply from the node. If the node gives back reply, it indicates that the node is not an attacked node but rather a node, which sends packet at a lower rate. However, if there is no reply from the node, the root then sends message again for the second and third time to the node because the first message might not have reached the node due to the unreliable nature of the network. However, if there is still no reply from the node, then it means than the node is attacked node and there is a blackhole node in the network.

3.3 Blackhole node identification

Once the existence of blackhole node is identified, the root then searches for the blackhole node. It starts searching from the parent of the attacked node and traverses through its ancestors towards the root. The first obtained non-delayed node or node with lower sequence delay percentage is the blackhole node. This is because in a blackhole attack, the blackhole node doesn't transfer packets of its children but its packets are transferred. percentage is the blackhole node. This is because in a blackhole attack, the blackhole node doesn't transfer packets of its children but its packets are transferred.

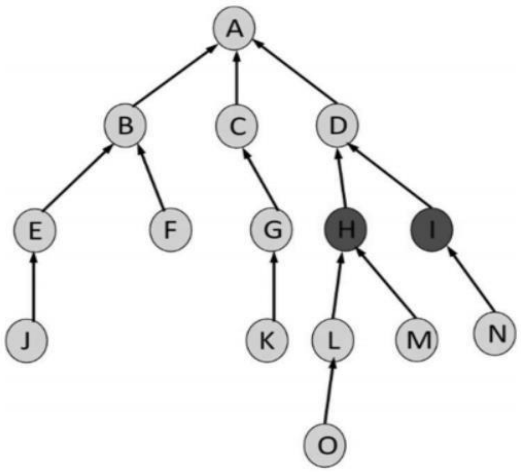


Figure 1: Multiple blackhole nodes

In figure 1, node A is the root node and node H & node I are blackhole nodes whereas node L, M, O & N are victim nodes. When one or all of node L, M or O are identified as attacked nodes, the root looks for a node with a non-delayed sequence delay percentage in their ancestor. When it reaches node H, it will identify it as a blackhole node and stops searching since Node H is not delayed. Similarly, after node N is identified as an attacked node, the root search for node N's ancestors and will find node I which is the parent of node N and a non-delayed node. Hence, node H and node I will correctly be identified as blackhole nodes.

3.4 Blackhole node removal

After the blackhole node is identified, the root sends message to the attacked nodes through their neighbours. When the blackhole node's children receives the message, they will drop the blackhole node from their parent list and will choose a node among their

neighbors, as their new parent. The blackhole node's neighbours will also remove it from their neighbours list. Consequently, the blackhole node will be isolated and removed from the network.

4. Experimentation

In order to evaluate the proposed algorithm, blackhole attack is simulated using Cooja simulator. Cooja is the network simulator for Contiki which is an operating systems used for wireless sensor networks. Two networks with different sizes are simulated and blackhole attack is applied in the networks. In the smaller network, a total of 12 nodes which contains 1 sink node, 2 blackhole nodes and 9 sender nodes is used. The larger network contains a total of 22 nodes with 1 sink node, 3 blackhole nodes and the remaining 18 sender nodes. The simulation is executed in both networks with and without the proposed algorithm which is a total of four simulations.

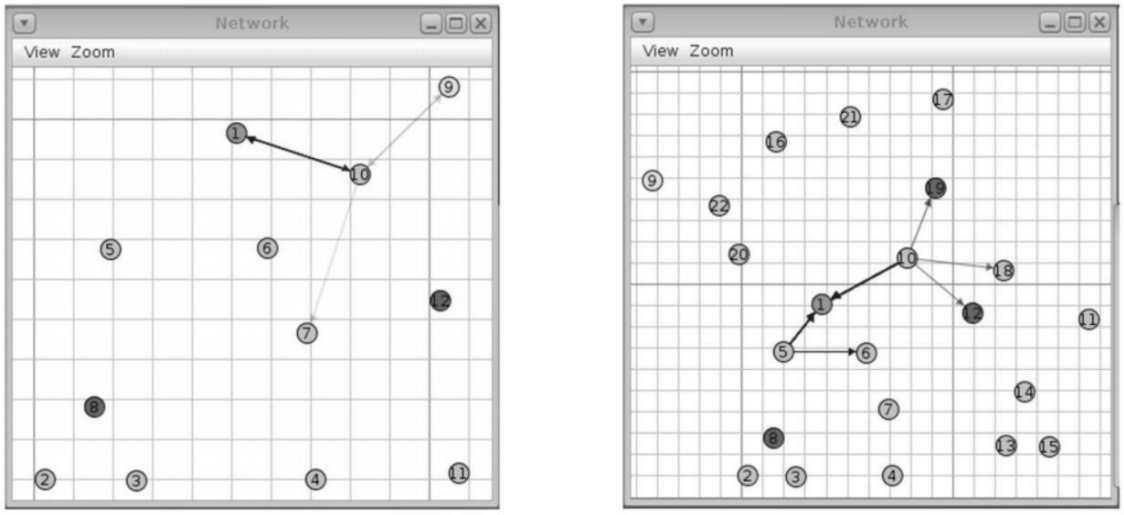


Figure 2: Simulation interface for smaller and larger networks

Figure 2, shows the simulation interface for the smaller and larger networks. In the smaller network, node 1 is root node, node 8 & 12 are blackhole nodes and the other nodes are legitimate nodes. In the larger network, node 1 is the root node, node 8, 12 & 19 are blackhole nodes and the other nodes are legitimate nodes. By applying the proposed algorithm, blackhole attack is detected and the blackhole nodes are identified. The attacked nodes have chosen a new parent and their packets were able to reach the root afterwards.Using the provided algorithm, packet delivery rate of the smaller network has improved by 42.1% and the packet delivery rate of the larger network has improved by 31.2%. This shows that packets which were previously being dropped by the blackhole nodes, are able to reach their destination when the provided algorithm is applied. There is an increase in energy consumption of the smaller and larger networks by 4.6% and 6.2% respectively using the proposed algorithm which is not high. This is important since IoT contains resource constrained devices.

5. Conclusions

In this work, an algorithm have been proposed to mitigate blackhole attack on the routing protocol RPL. The algorithm detects the presence of blackhole attack, identifies the attacker node and removes the attacker node from the network. The provided algorithm is tested by simulating blackhole attack. The result obtained in the simulation shows an improvement in packet delivery rate. This shows that the provided algorithm has succeed in detecting blackhole attack and removing attacker node. Furthermore, the energy overhead due to the provided algorithm is within acceptable limits. This shows that the proposed algorithm has provento be effective for detection of blackhole attack on RPL.

The proposed algorithm in this work has been tested and been proven to be viable for a network containing up to 22 nodes. For future work, providing a solution which could be viable for larger network size is recommended. Additionally, this work focuses only on blackhole attack. However, considering different attacks occurring at the same time could give the work more depth.

References

[1] Meron Mekonnen, "An Algorithm for Detection of Blackhole Attack on RPL," *A thesis submitted in partial fulfillment of the requirements for the Degree of Master of Science in Software Engineering, HiLCoE School of Computer Science and Technology,* December 2022.

[2] J. Vasseur, N. Agarwal, J. Hui, Z. Shelby, P. Bertrand and C. Chauvenet, "RPL: The IP routing protocol designed for low power and lossy networks," *Internet Protocol for Smart Objects (IPSO) Alliance,* April 2011.

[3] T. Tsvetkov and A. Klein, "RPL: IPv6 routing protocol for low power and lossy networks," *Network,* July 2011.

[4] J. V. V. Sobra, J. J. P. C. Rodrigues, R. A. L. Rabêlo, J. Al-Muhtadi and V. Korotaev, "Routing Protocols for Low Power and Lossy Networks in Internet of Things Applications," *Sensors,* vol. 19, no. 9, January 2019.

[5] P. Pongle and G. Chavan, "A survey: Attacks on RPL and 6LoWPAN in IoT," *In2015 International conference on pervasive computing (ICPC),* pp. 1-6, January 2015

[6] F. Ahmed and Y.-B. Ko, "Mitigation of black hole attacks in Routing Protocol for Low Power and Lossy Networks," *Security and Communication Networks,* vol. 9, pp. 5143-5154, October 2016

[7] J. Jiang, Y. Liu and B. Dezfouli, "A Root-based Defense Mechanism Against RPL Blackhole Attacks in Internet of Things Networks," *In 2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), IEEE,* pp. 1194-1199, 2018

Face Orientation Quantification for Face Sequence Indexing and Matching: The Case of Smartphone

Natnael Dereje Estefanos
HiLCoE School of Computer Science and Technology, Addis Ababa, Ethiopia
natnael.dereje27@gmail.com,

Fekade Getahun
Addis Ababa University, Addis Ababa, Ethiopia
fekadegetahun@gmail.com

Abstract

In recent years, there has been a lot of research into Face Orientation Quantification for Face Sequence Indexing and Matching. There are two primary reasons to write this thesis paper: the first is to provide an up-to-date analysis of the current literature; the second is to reveal some technical terms that describe Face Orientation Quantification for Face Sequence Indexing and Matching. Face Orientation Quantification is a well-developed topic of software engineering, so it is no wonder that we were inundated with academic articles. We tried to review several such articles and initiatives, many of which we found immediately applicable in our thesis of interest. The use of Face orientation Quantification in social life indicates the embrace of recreational principles by its detractors, as well as spontaneity. Face Orientation Quantification enables the emergence of new categories; these are essentially matrices of complicated behaviors, but during the Quantification process, they are transformed into measurable value.

However, in domains such as close relationships and identities, these attempts frequently change the nature of the category and disappear. Simultaneously, the dominant Quantification obliterates existing objects and relationships, leaving many social processes fundamentally invisible and unmeasurable. Finally, orientation Quantification contends that it is commonly addressed in sectors where arithmetical significance is lacking.

Keywords*: Face Orientation Quantification, Face Sequence Indexing and Matching, Local Binary Pattern, Principal Component Analysis.*

1 Introduction

Communication and information technologies are becoming increasingly ubiquitous in many parts of our lives and enterprises. Ushering in a world of previously unseen scenarios in which People engage with technological gadgets implanted in sensitive surroundings, to and sensitive to their existence Indeed, since the first instances of "smart" buildings with computer-assisted security and fire safety mechanisms were introduced, demotic research has been characterized by a demand for more sophisticated services tailored to each user's specific needs. The role of person sensing and recognition becomes of fundamental importance in an ecosystem regarded as a "community" of

smart objects driven by computational capabilities and high user-friendliness, capable of detecting and responding to the presence of various persons in a smooth, non-intrusive, and frequently undetectable manner.

Face Orientation Quantification has become one of the most important smartphone vision challenges explored in the literature thus far. In / out rotations are frequently generated by natural head movement. When the face is rotated along the axis from the face to the observer, in-plane movement is observed (rotation with respect to z axis), applying face Orientation Quantification for face sequence indexing and Face matching is an essential step in the development of automatic face analysis research.

Face Orientation Quantification for Face Sequence Indexing and Matching is an issue that has been discussed It has been around for almost 20 years, but it is still unsolved in terms of speed and accuracy. A conventional face detection method splits an image into tiny sub-windows and use a qualified classifier to determine whether or not a face is there. or not each sub-window contains a face image. There has been a rapid development of predicting the face image size of the sub-windows in the last decade.

The traditional Face Orientation Quantification (FOQ) can be categorized into two categories: holistic features and local feature approaches. The holistic category is further subdivided into linear and nonlinear prediction techniques. Since not all of the faces in an image are the same size, predicting the size of the sub-windows is difficult. As a result, the Face Orientation Quantification step must be replicated several times at various resolution levels. Color models are used in many Face Sequence Indexing and Matching (FSIM) projects.

According to the values of the pixels concerning the image pixels can be classified as skin or non-skin pixels. Although color diversity is a valuable indication of the existence of faces in photos, it should not be relied on alone to locate them.

Smartphones have evolved into the most simple and pervasive interface for accessing Internet services. They come with a variety of sensors that can greatly increase the user's cognitive abilities. Thanks to the broad variety of possible applications, face recognition is one of the fastest growing applications in the smartphone domain. Face Orientation Quantification, on the other hand, is a computationally intensive method that necessitates resources that are beyond the capabilities of even modern mobile devices.

2 Related Works

Face Orientation Quantification issues have been the subject of several studies [16]. Face Sequence Indexing and Matching, on the other hand, has received little attention. Most currently available face recognition approaches run into the orientation problem [14], which makes neutral faces difficult to be identified when photographs of them appear with various facial expressions. Some previous studies have attempted to predict the success or failure of the Face Orientation Quantification verification result, which is akin to the pursuit of a verification method.

[21] identify the most recent findings and study patterns in their survey on the state of the art in multi-modal face recognition, noting that "the variety and complexity of algorithmic techniques investigated is growing." Enhancing identification accuracy, increasing resilience to facial expressions, and, more recently, improving algorithm performance are the main problems in this field.

2.1 Projection Methods

The principal component analysis (PCA) which is a dimensionality reduction approach is one common methodology underlying several face recognition projection algorithms. The PCA technique traditionally employed by Turkey and Pentland in the Eigenfaces approach [6] calculates the own vectors and values of a covariance matrix and resolves a problem of a value of its own. This approach does not discriminate but instead the vectors that meet the highest individual values are utilized in a low-dimensional space as the basic vectors. Then the picture may be projected onto this smaller area and the closest match can be found via Euclidean distance. Also investigated was the merging of Eigenfaces and Eigen modules into a hybrid approach known as modular Eigenfaces. Results are coupled with proprieties, eyes and eyes which, in addition to their own surfaces, are computed in a similar way.

2.2 Statistical methods

This approach uses the breakdown of space in high-dimensional picture areas to assess the full density functions. The matching is then made in the maximum probability formulation utilizing these density estimates rather than using the normal distance metric. The Independent Component Analysis (ICA), which seeks to calculate separate properties of things such as the human face, is another approach comparable to PCA.

Hidden Markov models are the most popular statistical method based on features (HMM). It is based on a Markov chain with a limited number of states, a State transition probability matrix, a distribution of the starting state probability and a collection of probability density functions for every state. The first thing is to scan pictures in 1D or 2D and extract a collection of blocks. Features are retrieved from top to bottom as they are naturally present, and a status is assigned to each face area in the HMM. Finally, the most likely match is to be found with a maximum selection process. The computational complexity of earlier techniques of this was significantly reduced by [20]

2.3 Graph matching methods

Elastic Graph Matching (EGM) is a genetic algorithms object identification method used in computer vision. It is biologically inspired by two sources: (I) the visual characteristics utilized are based on Gabor wavelets, which have been shown to be a suitable model of early visual information in the brain, namely simple cells in the primary visual cortex. (ii) The matching algorithm is an algorithmic version of dynamic link matching (DLM).

EGM represents visual objects as labeled networks, with nodes representing local textures based on Gabor wavelets and edges representing distances between node positions on an image. As a result, a picture of an item is represented as a collection of local textures arranged in a certain spatial arrangement. When a new item in an image has to be identified, the labeled graphs of previously recorded objects, known as model graphs, are matched onto the picture. The locations of the nodes in the picture for each model graph are adjusted such that the local texture of the image matches the local surface of the model and the distances between the locations match the distances between the nodes of the model.

2.4 Neural network methods

Neural networks such as the probabilistic- based neural network (PDBNN) and evolving pursuit Expression Point (EP) were also explored for greater generalization in face classes . These kinds of learning methods enhance the decision-making line using fitness functions. Rowley H., Baluja. [3] deployed the PDBNN, which consisted of three modules: face detector, eye locator and face detector. It employs the hierarchical network structure with non-linear effects in order to recognize the intensity of features and the edge information.

Neural networks provide data processing similar to that seen in biological systems such as the human brain, where information is processed. Its key strengths are their ability to learn from examples, their tolerance for failure, and their strength.

They may be used to recognize a wide range of face images while requiring little modification to the recognition algorithm.

3 Face Sequence Extraction

A series of faces corresponding to the picture is retrieved for the face image. The size of the extracted face in the image is a significant consideration in this situation. In general, we utilize a basic tracking approach between adjacent Index-frames in some of the collected face images, which checks the overlapping of the bounding boxes of the retrieved faces. If the faces are almost in the same position and more than 30% of the area overlaps, they are thought to belong to the same individual. Face detection is applied first to every frame within a specific interval of frames. The system employs a neural network-based face detector that identifies primarily frontal faces of varied sizes and positions.

Sequence indexing and matching performed with the nearest pair of the face sequence and class can be used to index and match the face. A facial class can be found inside a face class that was incorrectly classified and is notoriously far from the rest. This face may be incredibly close mistakenly to a sequence that does not meet the face class but is matched due of the incorrectly classified face. In this case, we may be able to steal the erroneous face by taking advantage of the fact that the wrong face does not constitute a statistical majority. This problem may thus be solved by extracting and using face-class prediction techniques.

Using facial feature extraction for face orientation quantification, the face sequence extraction approach could be greatly improved, and outliers could be avoided by using temporal redundancy

within subsequent picture frames. We are aware that our method is incapable of dealing with large changes in picture structure. since the underlying PCA is affected by scale, rotation, and changing illumination conditions Our preliminary face clustering attempts have been positive, and we want to examine the potential of enhancing our face orientation Quantification by modelling the distribution of face sequence indexing and matching.

4 Solution Design

A face is a typical non-rigid object, and its shape changes dramatically as a result of its expression, lighting, and occlusion. As a result, face orientation Quantification should be capable of reliably drawing out the distinguishing feature from facial pictures with varied disruptions. The outputs of face orientation are usually the face pictures to be accurately identified. In the face pictures, the orientation of the facial organs is generally stable and precise. As a result, we adopt a technique based on face orientation Quantification for face sequence indexing Matching.

4.1 Face orientation Quantification representation

Face Orientation Quantification may be used to index and match the face by applying it to the face sequence and class. A facial orientation can reflect a face class that was misidentified and is famously different from the others. This face might be mistakenly matched to a sequence that does not belong in the face class but is matched because of the improperly identified face. In this situation, we may be able to detect the incorrect face by using the fact that the incorrect face does not form a statistical majority. By extracting and employing face-class orientation Quantification, this problem may be overcome, methods for representation of face orientation Quantification to facial classes include defining the difference between different orientation angle using the face orientation model. The basic idea is taken from the face's closest pair-based sequence; nevertheless, this might be less sensitive to curves. This is due to the fact that the method relies on statistical models for face classification and is expected to be less effective for groups with statistically tiny outliers

4.2 Face orientation Quantification model

A Face sequence indexing and matching based on face orientation Quantification is the core module in this design. On this section, we've summarized the key design considerations. A distance metric will be used in order to estimate the distance between the trained face orientation and the proposed Quantification. The face image is finally given with a list of the shortest distances. A face orientation Quantification application requires the user to assert their identification

Preliminary Face alignment has few options, and the most of them are computationally difficult to accomplish. Even if it isn't centered, rotated, or scaled properly all at once. You might be able to identify a common strategy to align them by using a variety of facial images. Each pixel stack in a set is assigned a probability value, and a series of translations, rotations, and scaling are iterated to discover the best transformation that minimizes the difference between a set of face images. The alignment is accomplished by transforming each image one at a time while reducing the total entropy sum.

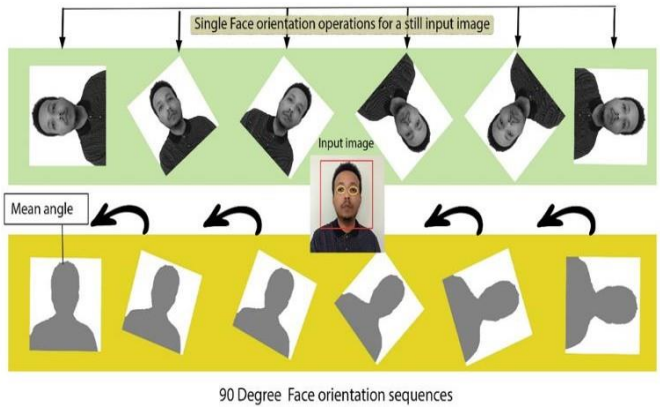


***Figure 1**, A processing diagram demonstrating*

4.3 PCA and Cascade classifier

A feature-based method is used to train the cascade classifiers. The cascade classifier is made up of a series of tests on input properties. Selected attributes are divided into stages, with each step being trained as a strong classifier using the best weak classifiers. For weak classifiers, the tested approach uses a high number of LBP.

Each stage is in charge of determining whether or not the detection window contains a face. If the window fails at any point, it will be deleted immediately. As a consequence of the cascading, regions without faces will be removed at each level, allowing for speedier processing. During the learning process, the number of phases is determined and set in order to reach a certain detection accuracy.

From face photographs, the algorithms extract a set of geometrical attributes and distances that can then be compared. A feature-based algorithm like local feature analysis is an example. Methods for Measuring Face Orientation.

4.4 Proposed System Architecture

Face Orientation Quantification has adopted state-of-the-art techniques from domains including different orientation factor processing, computer vision, pattern recognition, and machine learning, among others. All approaches have their own pros and limitations, making it difficult to choose the optimal one for a given task. To perform the Face Orientation Quantification, efficiency and low complexity are key considerations.

Feature-based approaches have made significant steps in our study, and holistic methods, it is believed, still perform better overall. To normalize orientation Quantification, however, facial characteristics must be detected and used. Performance is also affected by variations in scale, head posture, and partial occlusion. However, most of these approaches are partially adaptable enough to function in all situations, while others are too sophisticated or inefficient for the case of smartphone. it is possible to process the database pictures such that the training data is at least standardized.

This might result in a obtaining of discriminant information when matching a new Face,

depending on the Face Orientation Quantification technique. we cannot assume that the training set would include every conceivable variation that might appear in the Quantification Face pictures. The precise set of potential modifications would be helpful, however that depends on the user and varies from camera to camera. When estimating approaches, the major concern was that the algorithms would operate within the constraints of smartphone.

Our knowledge of the Face orientation Quantification is challenged by the extraordinary effectiveness of human face recognition.

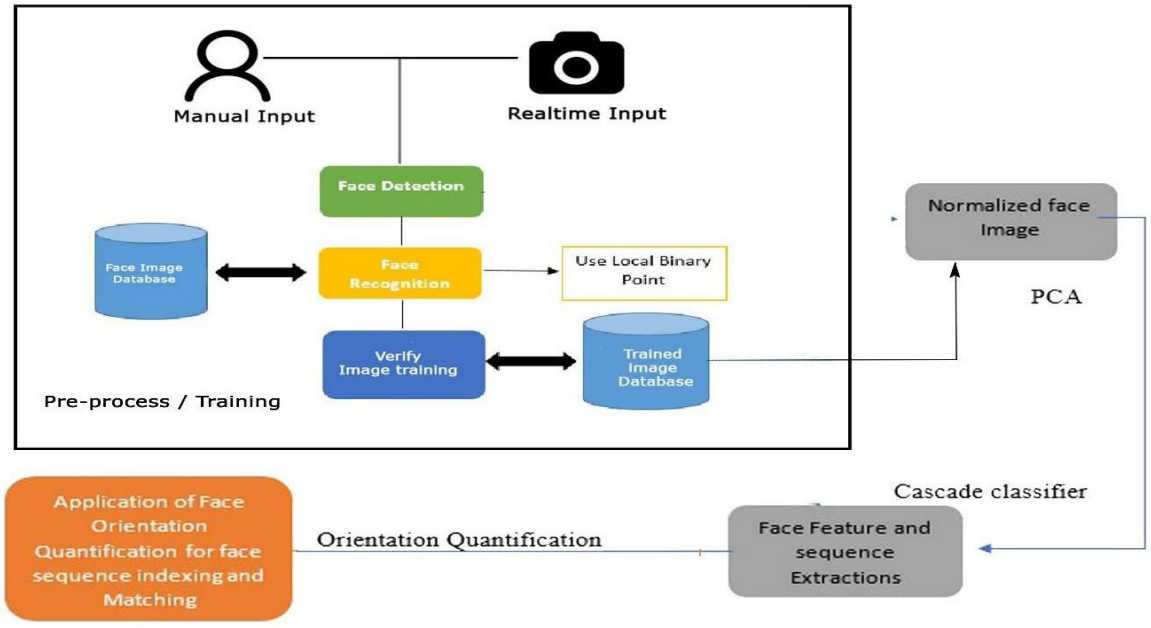


***Figure 2**, Face Orientation Quantification Architecture Model*

And our efficiency comes from observer- driven biases at high-level levels of orientation Quantification processing that favor upright faces' higher-order characteristics features like eyes, nose, and mouth and their orientation spatial arrangement within the face.

Face orientation Quantification in the face sequence indexing and matching with the encoding of local edge orientation in the primary visual cortex. Face image pattern responses generally indicate a preference orientation with a peak level of response, with diminishing responsiveness for edges oriented away from the desired orientation. The function that connects face orientation Quantification responses to direction is a sequence indexing with a bandwidth that represents the sharpness of the face orientation.

5 Implementation and Evaluation

There are few algorithms for face alignment available, and the most of them are computationally costly to implement. Even if it is not properly cantered, rotated, and enlarged at the same time.

Using a variety of facial images, you may discover a common strategy for aligning them. Iteratively apply a set of translations, rotations, and scaling to discover the best transformation that minimizes the difference between a set of face images. Each pixel stack in the set is assigned a probability value. The alignment is accomplished by transforming each image one at a time while reducing the total entropy sum.

We must first authenticate a face picture from a captured image before we can employ face orientation Quantification to face sequence indexing. Then start adjusting the face's orientation. One way is to use Eigen to compare selected facial picture Principal Component Analysis (PCA). The simplest and quickest algorithm is PCA. PCA is a technique for turning a set of possibly associated parameters in a sample face image into a set of uncorrelated variables known as principal components. The primary goal of PCA is to find correlations in data.

The change in the values of two or more variables at the same time is quantified by correlation. In this instance, the linear correlation is used. Consider our dataset, which comprises n parametric variables retrieved from the observations (x1,x2,...,xn). PCA's purpose is to choose k n (typically k = 2 or 3) new variables (that will turn out to be the main components) that characterize a significant portion of the orientation Quantification captured in the face picture by accounting for the biggest covariations imaginable in it. We'll look at n = 3 variables obtained from m = 9 different orientation angles to demonstrate the manner of representation. the initial variables and the face-orientation technique Face Orientation Quantification for Indexing and Matching of Face Sequences

5.1 Analysis and Image processing filters

The Quantification 's flexibility to data collection scenarios is supported by the underlying fundamental assumption, which is supported by the artefact and design methods, as well as Orientation Map Matching Orientation and Map Extraction pixel-based edge orientation. This is to collect data on Face Orientation Quantification from multiple sources using Face Sequence Indexing and Matching.

The most essential techniques to illumination normalization are what are known as quasi illumination-invariant image filters. High- pass and locally scaled high-pass filters are examples of these, which are frequently based on fairly simple picture creation models, such as modeling light as a spatially low-frequency band of the Fourier spectrum and identity-based information as a high-frequency band.

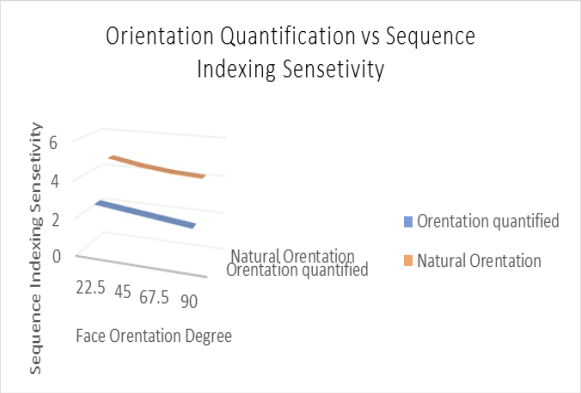


Figure 3, *Line graphs depict the Orientation Quantification versus Sequence Indexing Sensitivity*

The amplitude values at all spatial frequencies inside each of these orientation ranges were added together and divided by the overall maximum total for each picture. Face picture stimuli is being used Greyscale face photos were filtered in 22.5° increments from 0° to 90° vertical orientations and back. The outlines of filtered faces were examined using face orientation quantifications or natural orientation and describes the hypothesized pattern of human face index sensitivity to face orientation quantifications as a function of orientation range, for both vertical and inverted faces in the imageplane

5.2 Testing and Evaluation

Throughout the image capture process, face photos are cropped and changed to grayscale images, which are then saved in the same folder to construct face databases for extraction operations. Following that, the standardization technique is used to all photographs in order to minimize noise and establish the correct image scaling position in order to obtain the recognition result as quickly as possible.

A face sequence indexing application based on face orientation measurement is envisioned in this technique. We've put up a list of the most pressing design issues in this area. A distance meter will be used to determine the distance between the taught face orientation and the suggested quantification. Finally, a list of the shortest distances is included with the face image. The user must identify their identification in order for a face orientation Quantification system to work. A threshold is typically used to determine if the distance between the input face and the face in the database is near enough. As a result, prior to authentication, Face Orientation Quantification is performed.

Consider a dataset with m observations and n parametric variables (x1,x2,...,xn). The purpose of PCA is to identify new variables (that will turn out to be the primary components) that decide a large percentage of the information stored in it by accounting for the maximum covariations conceivable in the data.

5.3 Performance and speed results

We acknowledge in our study that the LBP face recognition algorithm's capacity is largely dependent on the accuracy of the feature extraction and comparison stages, which is also highly dependent on the quality of both the input and training/reference pictures used in the face comparison phase. To increase the LBP algorithm's face recognition accuracy, lighting, sharpness, noise, resolution, scale, and posture were used to create the best match pictures, which revealed more details of image characteristicsfor more accurate feature extraction and comparison.

6 Results and Discussion

The majority of the methods indicated in the literature review essentially use many face orientation strategies in parallel, each specialized to one view and based on the PCA methodology described above for face orientation of the same views. The subsystem that is most equipped to describe the image data in terms of its orientation scale is generally the one that is specialized for a perspective that is closest to the view of the facial picture. As a result, it may be chosen automatically to execute the face orientation quantification. The face sequence and indexing show how well a system can perform if it does not adjust for the impacts of rotation in depth and instead depends solely on the robustness of the subsystem nearest to the look at the probe image Because matching is done with a local binary pattern, our fundamental orientation quantification only adjusts for rotation in depth.

Before we can use Face Orientation Quantification for face sequence indexing and matching, we must first prepare the face dataset for training. The face dataset was created using face recognition technology, which identifies and records faces in real- time cameras. The images are kept in a dataset folder and may be used in the future for face recognition, feature extraction, and training. The gadget will first collect information about the user, such as eye, noise, and mouth movements, before opening the camera and photographing the person in 42 different face expressions and moods. The images are saved in a dataset folder with the same unique ID as the individual's information, which is stored in a database.

	Number of Cycles (x 106)	Computation Time (With a CPU at 500Mhz)	Performance Memory Consumption for Data
Face detection	1116	1.92 sec	~ 442 KB
Eye localization	475	1.17 sec	~ 336 KB
Face normalization	46	1.11 sec	~ 112 KB
Face classification (PCA & LBP)	21	0.12 sec	~ 253 KB
Face classification	1	0.11 sec	~ 1044 KB

Table 1 *Performance and speed result*

The Face orientation measurement regarding community participation reveals the opponents' support of leisure ideals and spontaneity. The facial expression Quantification enables the creation of new categories; these are essentially matrices of complicated behaviors that are transformed into 'things' via the Quantification process. These attempts, on the other hand, usually have an impact on the substance of the category and vanish in domains such as close connections and identities. At the same time, the dominant Quantification obliterates existing objects and relationships, leaving many social processes invisible and unmeasurable. Lastly, critics of orientation Quantification claim that it is widely used in businesses where statistical significance is absent. Despite these difficulties, face orientation quantification is an inescapable technique

7 Conclusion and Future Work

We utilized the Local Binary Patterns histogram technique to recognize faces in the proposed Face Orientation Quantification for face sequence indexing and matching. Image-based face identification, facial feature extraction, and image-based sequence index classification are the three aspects of the method. Face Orientation Quantification is a technique for determining the orientation of a person's face in a picture. In the feature extraction phase, facial landmarks are detected and used to create an LBP histogram that delivers a totally unique output, and in the identification phase, the input image's histogram is compared to the database histogram using a classifier. The results demonstrate that the system can distinguish between known and unknown individuals.

In real-time, our algorithms can identify and recognize faces with great accuracy. When compared to other detection methods, it has a faster detection speed. The identification of eyes is used to improve face detection accuracy. Face component alignment, picture smoothing, and contrast modulation all help to improve face detection performance. Face images are taken in real time as training samples and identified in a variety of situations, including amid other faces. New classifiers will be learned in the future to broaden face recognition across a wider variety of facial positions. In order to fine- tune the face picture and sustain effective evaluation, the system may foresee head rotation.

Finally, while some attempts have been made to deal with face Orientation quantitative processes, other forms of Face orientation quantitative applications necessitate the use of specific digital approaches in the most efficient manner possible. While certain approaches have been addressed in research, additional effort must be done to enhance face orientation quantification techniques that accommodates the image factors mention in the scope limitation of theses study includes such as image-specific issues, or subject-specific factors such as, image demographics and frontmost part of the eye such as the cornea furthermore, it is suggested that still there remains a gap in terms of study in video image recognition that requires Similarly to the face image orientation quantification, which is more complicated that needs to be researched

References

[1] T. Sim, S. Baker, and M. Bsat. The CMU Pose, lumination, and Expression (PIE) database of human faces. Technical Report CMU-RI-TR-01-02, Robotics Institute, Carnegie Mellon University, Pittsburgh, PA,January 2001.

[2] Kelly M. D., Visual Identification of People by Computer. Stanford AI Project, Stanford CA, Technical Report, 1970,

[3] Rowley H., Baluja S., and Kanade T., Neural network-based face detection, IEEE Trans. Pattern Anal. Mach. Intell, 20, 1998,

[4] T. Fromherz, P. Stucki, M. Bichsel, "A survey of face recognition," MML Technical Report, No 97.01, Dept. of Computer Science, University of Zurich, Zurich, 1997.

[5] T. Riklin-Raviv and A. Shashua, "The Quotient image: Class based recognition and synthesis under varying illumination conditions," In CVPR, P. II

[6] Turk, M. A., Pentland, A. P., Eigenfaces for recognition, 1991, Cognitive Neurosci., V.

[7] K. Jonsson, J. Matas, J. Kittler, and Y. Li, "Learning support vectors for face verification and recognition," In Proc. IEEE International Conference on Automatic Face and Gesture Recognition, 2000.

[8] R. Brunelli and T. Poggio, "Face recognition: Features versus templates," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 15, no. 10, pp. 1002–1112, 1993.

[9] C. Schneider, N. Esau, L. Kleinjohann, and B. Kleinjohann 2006. Feature based face localization and recognition on mobile devices. Control, Automation, Robotics and Vision,

[10] Lu, H. et.al. (2008). Mpca: Multilinear principal component analysis of tensor objects. IEEE Transactions on Neural Networks, 19.

[11] Wang, J. et.al. (2010). Multilinear principal component analysis for face recognition with fewer features. Neurocomputing, Elsevier

[12] S. Lawrence, C. L. Giles, A. C. Tsoi and

[13] A. D. Back, "Face recognition: a convolutional neural network approach," IEEE Transactions on Neural Network, vol. 8, no. 1, Jan 1997.

[14] M. Black and Y. Yacoob. Tracking and recognizing rigid and non-rigid facial motions using local parametric models of image motion. In Proc. of International conference on Computer Vision, pages 354– 381, 1995.

[15] P. Viola and M.J. Jones, "Rapid object detection using a boosted cascade of simple Features," IEEE Transactions on Computer Vision and Pattern Recognition, vol. 1, 2001.

[16] Kelly M. D., Visual Identification of People by Computer. Stanford AI Project, Stanford, CA, Technical Report, 1970, AI- 130

[17] A. Lanitis, C. Taylor, and T. Cootes, "Automatic interpretation and coding of face images using flexible models," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 19, no. 7, 1997.

[18] Pentland A. and Moghaddam B., Probabilistic Visual Learning for Object Representation, IEEE Trans. Pattern Analysis and Machine Intelligence,19(7), July 1997,

[19] Rapha el Feraud, Olivier J. Bernier, Jean-Emmanuel Viallet, and Michel Collobert, A Fast and Accurate Face Detector Based on Neural Networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 23, No. 1, January 2001.

[20] Wiskott L., Fellous J.-M., and Von dermalsburg C., (1997), Face recognition by elastic bunch graph matching. IEEE Trans. Patt. Anal. Mach. Intell. 19, (1997) 725-779..

[21] Fasel B., and Luettin J., "Automatic facial expression analysis: A survey," Pattern Recognit.,36(1), 2003.

[22] Kittler J., Hatef M., Duin R.P.W., and Matas J., On combining classifiers. IEEE Trans. Pattern Analysis and Machine Intelligence, 20(3), 1998,

[23] Bowyer, K.; Chang, K. & Flynn, P. (2004). A survey of 3D and multi- modal 3D+2D face recognition, in Proc. Int. Conf. on Pattern Recognition (ICPR).

Spatial Cloud Data Management Framework for Emergency Management

Tesfamariam Damte
HiLCoE Software Engineering Programee, Ethiopia
tesfadamte@gmail.com

Fekade Getahun
HiLCoE, Ethiopia Department of Computer Science**,** Addis Ababa University, Ethiopia
fekadegetahun@gmail.com

Abstract

Effective management of spatial data is critical for emergency management, and the availability of spatial data through cloud computing demands an efficient framework. This research aims to develop a spatial cloud data management framework for emergency management, capable of handling and analyzing large volumes of spatial data. Objectives include exploring spatial data types, designing the framework with key features, and evaluating its performance during emergencies.

Following a design science approach, the framework comprises four layers: infrastructure, data, application, and client. It leverages cloud storage, PostgreSQL databases, open-source spatial data sources, Apache Spark, PostGIS, and forecasting models. Supporting thick and thin cloud computing clients ensures accessibility for authorized users.

The significance of this study is evident in its transformative impact on emergency management through the integration of Spatial Big data and advanced analytics. By enhancing preparedness, response, and recovery, it contributes to minimizing the impact of disasters and protecting lives. The evaluation findings underscore its practical effectiveness, rendering it a valuable asset for both practitioners and stakeholders in the field.

In conclusion, the spatial cloud data management framework offers a viable solution for efficient spatial data management during emergencies. Its multi-layered architecture and support for cloud clients improve emergency management operations, with practical implications for the field and future advancements.

Keywords: Cloud storage, PostgreSQL, PostGIS, Apache Spark,

1. Introduction

Most of the things in this world are in dramatic change. Emergency events happen everywhere in every corner of the world. Emergency events imply all events that endanger normal functioning of services and companies, endanger lives or resources (living environment) as well as events that are threatening stability of state [1]. To handle such problems emergency management is required.

Emergency events are spatial in nature and geo spatial technology better suited to address them. However, in emergency management system Geo-spatial science faces great challenge in terms of data intensity, computing intensity, concurrency intensity and spatiotemporal intensity [2].This challenge cannot be address by a traditional desktop GIS and the use of new technologies is required.

The emerging cloud computing paradigm brings remarkable solution in challenges of spatial technology with elastic and on-demand access to massively pooled, insatiable, and affordable computing resources [2].

This research tries to design the architectural framework of spatial cloud data management in emergency management. Also address challenges of big spatial data.

2. Related Literature

In the study by X. Ge and H. Wang [3], the potential benefits of cloud-based services for managing extensive geographic data in emergency management are explored. Their approach involves a methodology to handle and analyze vast spatial data volumes during disaster management using cloud computing. Within this context, Aly et al [1] propose a GIS-based cloud computing model for emergency systems, envisioning its potential to enhance the efficiency of disaster management. This model suggests merging Geographic Information Systems (GIS) with cloud computing to create an effective and adaptable framework. Meanwhile, W.W. Song et al [4] propose a spatio-temporal cloud platform to support GIS applications, addressing challenges related to managing large spatio-temporal data and providing computational resources. This platform integrates distributed processing and cloud computing technologies, presenting a scalable and flexible solution. In the context of spatial data analytics, D. Jiang et al [5] emphasize the advantages of employing spatial data analytics for emergency response. They highlight interdisciplinary cooperation, real-time data processing, and visualization using advanced tools as key practices for implementing spatial big data analytics. Lastly, S. Gao et al [6] discuss the use of cloud computing for spatial data management, showcasing cloud-based GIS platforms, spatial data analytics tools, and visualization tools as existing solutions. Their paper stresses the significance of creating scalable, secure, and well-governed cloud-based systems while underscoring research gaps in the evaluation of cloud-based spatial data management solutions.

Overall, these studies collectively shed light on the potential of cloud-based services, GIS-cloud integration, spatio-temporal platforms, and spatial data analytics in enhancing emergency management. They underscore the need for scalable, secure, and adaptable systems that effectively handle diverse data sources and promote real-time decision-making during emergencies. However, gaps persist in efficient data compression, scalable algorithms, interdisciplinary cooperation, and comprehensive evaluation, leaving room for further research and innovation.

After reviewing related literature spatial cloud data management framework for emergency management typically should consists of components, including data acquisition, data storage, data processing, and data visualization.

Different studies proposed different solution for forecasting drought. Demisse, G. B et al [7] introduces a novel spatial drought index (SDI) utilizing satellite-derived data, emphasizing its significance in drought- prone regions' decision-making. The proposed SDI combines the Standardized Precipitation Index (SPI) and the Normalized Difference Vegetation Index (NDVI) through a weighted sum, enabling a comprehensive assessment of drought's spatial variability. Similarly, the study Demisse, G. B et al [8] underscores the importance of sub-national drought monitoring for improved management strategies.

In the context of vegetation health and drought conditions, Huete et al [9] demonstrates NDVI's utility in Australia. Employing NDVI to gauge vegetation cover through satellite imagery, it contributes to understanding the impact of drought on vegetation. Additionally, Nigussie, A et al [10] reveals the enhanced accuracy of combining SPI, NDVI, and rainfall anomaly data for drought risk assessment. Collectively, the reviewed literature highlights the significance of integrating SPI, NDVI, and rainfall anomaly for a holistic understanding of drought conditions.

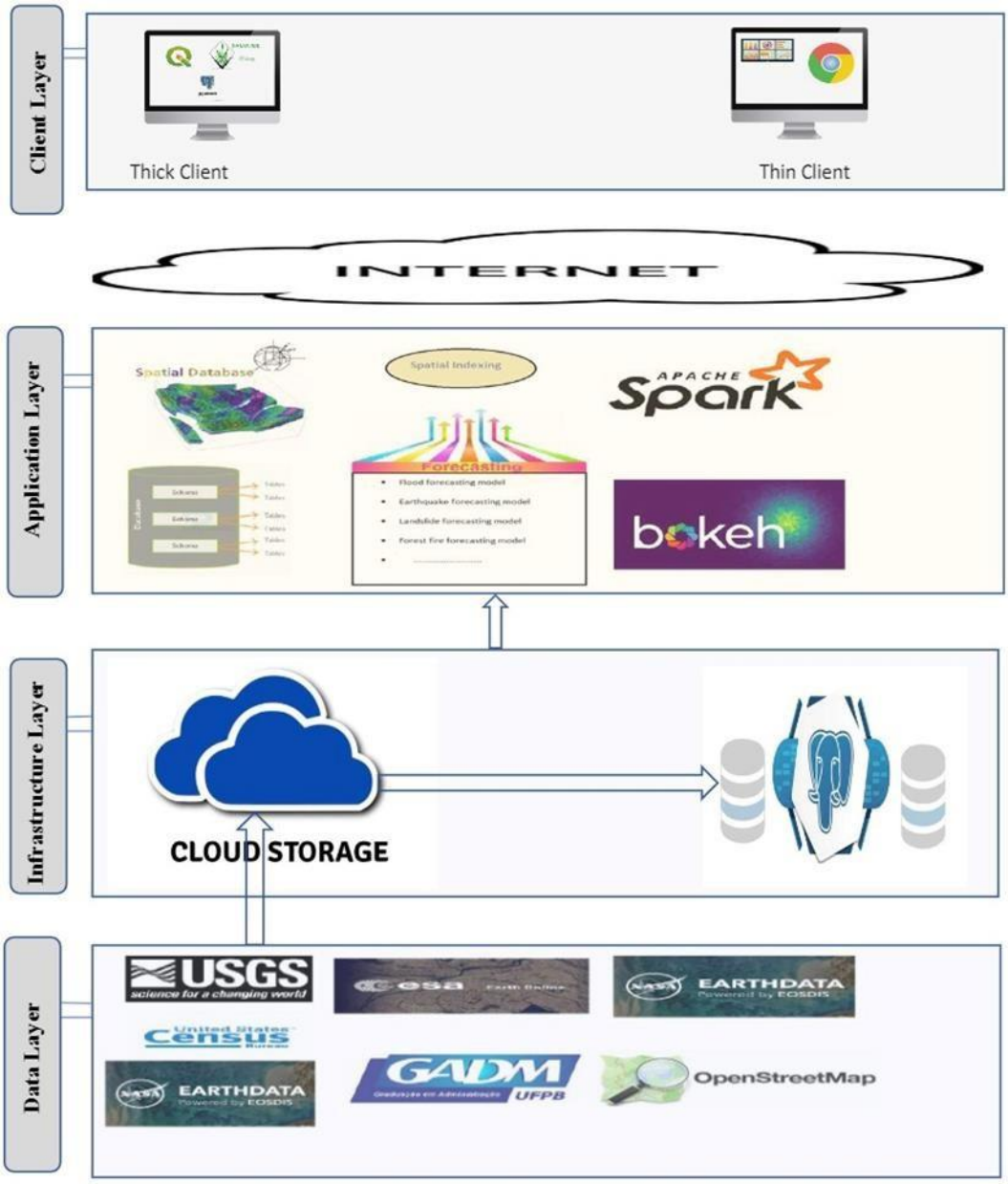
In conclusion, the literature highlights the significance of cloud-based spatial data management in emergency management, offering best practices and identifying research gaps. The use of SPI and NDVI

for drought forecasting models contributes to more accurate and comprehensive assessments of drought conditions.

3. FRAMEWORK OF SPATIAL CLOUD DATA MANAGEMENT FOR EMERGENCY MANAGEMENT

The proposed spatial cloud data management framework for emergency management consists of four layers: Data layer, Infrastructure layer, Application layer and Client layer.

Figure 1: A framework for spatial cloud data management for Emergency management.



3.1 Data Layer

The data layer serves as the foundation of the framework and incorporates various open-source spatial data providers, including USGS Earth Explorer, NASA Earth data, ESA Earth Online, USGS National Map Viewer, OpenStreetMap, GADM, US Census Bureau, CDO (climate data online), and Natural Earth. These sources offer access to a diverse range of satellite imagery, GIS data, demographic information,

climatedata, and more. To effectively integrate this data into a PostgreSQL database within the spatial cloud data management framework, a systematic approach is undertaken. The process involves identifying data sources, selecting a suitable cloud store service, creating storage containers, downloading data, transferring it to the cloud, and then configuring the PostgreSQL database to accept and organize the incoming data.

3.2 Infrastructure Layer

The infrastructure layer comprises two essential components: cloud storage and PostgreSQL database. Cloud storage options, including object storage, file storage, and block storage, offer scalability and flexibility for managing spatial data. Object storage is suitable for handling large volumes of unstructured data, while file storage is utilized for smaller datasets requiring file system access. Block storage provides persistent block-level storage volumes for spatial analysis on virtual machines. PostgreSQL, combined with the PostGIS spatial extension, emerges as a robust choice for spatial data management. It supports storage, indexing, and querying of vector and raster data, allowing for effective data organization and retrieval. The hybrid cloud deployment model is considered, incorporating both private and public cloud infrastructures for enhanced scalability and security.

3.3 Application Layer

The application layer encompasses key components, including spatial database, schema and table structures, spatial indexing, forecasting models, Apache Spark, and a web server. PostgreSQL with PostGIS supports data storage, retrieval, and manipulation, while schema and tables are defined to accommodate forecasting models. Spatial indexing enhances data querying efficiency, and Apache Spark facilitates real- time data streaming between sources and the PostgreSQL database. A drought forecasting model is demonstrated as a detailed example, incorporating NDVI calculations, rainfall analysis, SPI computation, and weighted averaging to assess drought risk. Finally, a web server, specifically Bokeh Server, is employed to create interactive web-based data visualizations and dashboards for users to analyze and comprehend the spatial data effectively.

3.4 Client Layer

The client layer encompasses both thin and thick clients. Thin clients, represented by web-based dashboards, offer lightweight access to the framework, enabling users to interact with the data easily. Thick clients, exemplified by QGIS software, provide more extensive capabilities for data analysis, manipulation, and visualization. QGIS is connected to the PostgreSQL database in the cloud, allowing users to create, edit, and analyze geospatial data in a secure and customized environment. This layered approach ensures accessibility, scalability, and adaptability for emergency management scenarios, empowering users to make informed decisions based on real-time spatial data.

The proposed spatial cloud data management framework integrates diverse data sources, cloud-based infrastructure, advanced analytics, and user-friendly interfaces to enhance emergency management processes. By leveraging cutting-edge technologies and methodologies, this framework offers a comprehensive solution for spatial data management, analysis, and visualization in time-sensitive situations.

4. Prototype

Emergency management teams benefit from cloud computing's ability to provide quick and secure access to data, aiding in rapid crisis response. The hybrid cloud deployment model is utilized in this

prototype, allowing spatial-based emergency management analysis through QGIS software. Direct connection to the cloud-based PostgreSQL database, facilitated by PostGIS, enables data analysis. Expertise in basic SQL queries and PostGIS functions is required for analysis on the private cloud. Flood forecasting involves processing Landsat ETM+ data, rainfall data, and SPI data from USGS and CDO, respectively. Spark streaming transports this data into the PostgreSQL database. Custom functions in PostgreSQL, such as NDVI and rainfall anomaly, transform Landsat ETM+ and rainfall data, enabling creation of NDVI and rainfall anomaly data. A customized drought risk function utilizes these data layers to generate a drought risk map, aiding in decision-making.

The custom function facilitates the calculation of drought risk based on the integration of various raster datasets. The function calculate_drought_risk is designed to process three essential input raster's: the Standard Precipitation Index (SPI) raster, rainfall anomaly raster, and Normalized Difference Vegetation Index (NDVI) raster. These input raster's are weighted according to their significance in drought risk assessment: SPI with a weight of 40%, rainfall anomaly with 30%, and NDVI with 30%.

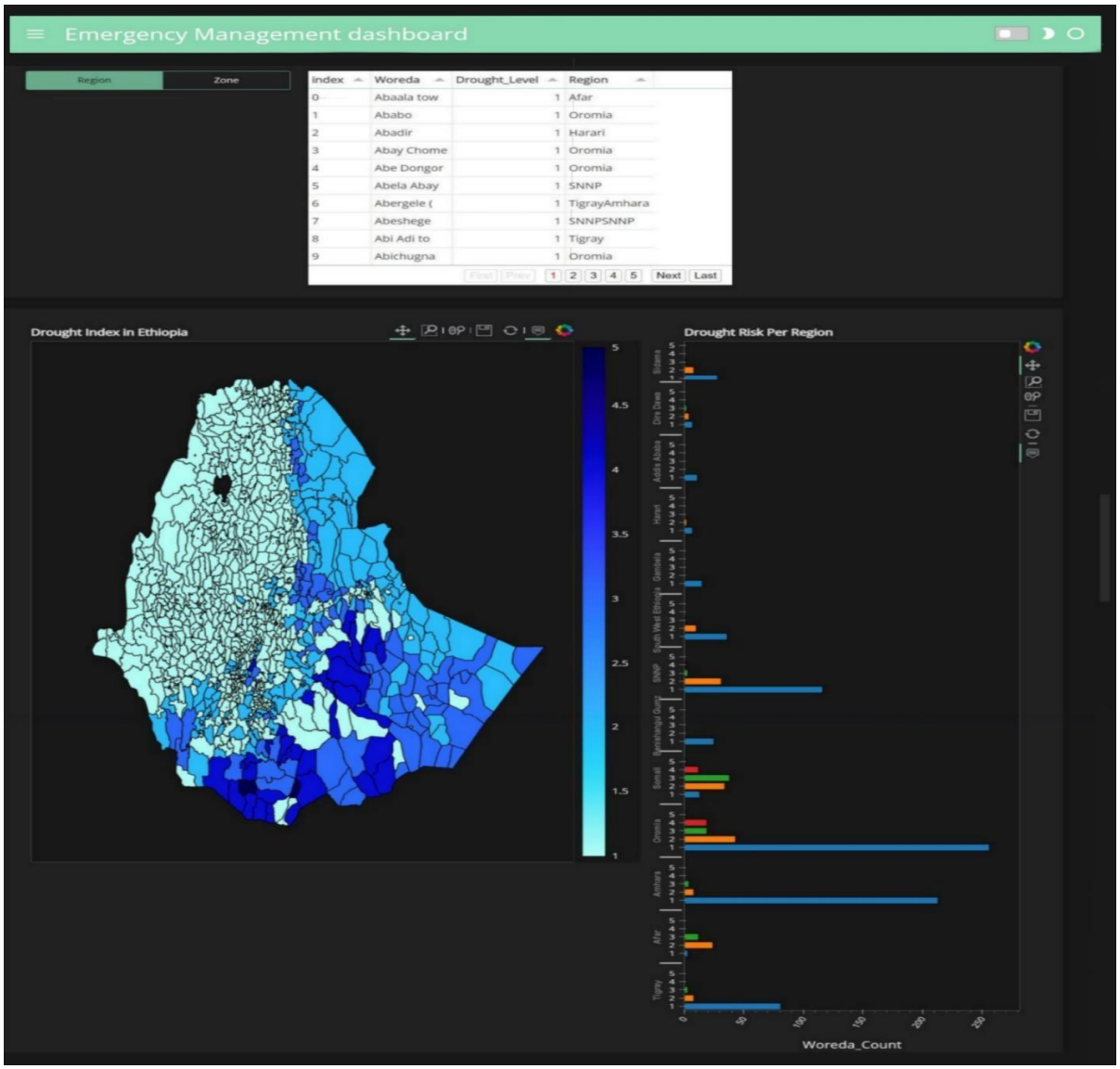
The algorithm dynamically creates or replaces a table named drought_map in the PostgreSQL database to store the resulting drought risk map. It begins by checking if the drought_map table already exists; if so, it clears the table, and if not, it creates the table with the necessary columns. The core of the algorithm involves performing arithmetic operations on the input rasters, applying the specified weights, and then combining them into an output raster representing the drought risk. This calculated output raster is then inserted into the drought_map table for visualization.

```
CREATE OR REPLACE FUNCTION calculate_drought_risk(spi_raster raster, rainfall_anomaly_raster raster, ndvi_raster raster, output_table_name text)
RETURNS void AS $$
DECLARE
   spi_weight float := 0.4;
   rainfall_anomaly_weight float := 0.3;
   ndvi_weight float := 0.3;
   output_raster raster;
BEGIN
   IF EXISTS (SELECT 1 FROM pg_tables WHERE tablename = output_table_name) THEN
      EXECUTE 'DELETE FROM ' || output_table_name;
   ELSE
      EXECUTE 'CREATE TABLE ' || output_table_name || ' (rid serial primary key, rast raster)';
   END IF;

   output_raster := (spi_raster * spi_weight + rainfall_anomaly_raster * rainfall_anomaly_weight + ndvi_raster * ndvi_weight)::raster;
   EXECUTE 'INSERT INTO ' || output_table_name || ' (rast) VALUES (' || quote_literal(output_raster) || ')';
END;
$$ LANGUAGE plpgsql;
```

To enhance data visualization, a web-based dashboard is crafted, utilizing Python libraries including panel, Hvplot, Geoview, Geopandas, and pandas. These open-source libraries create interactive plots, geospatial maps, and enable data retrieval from the PostgreSQL database. The dashboard serves as a comprehensive tool for decision-makers and the public, providing insights into emergency management. Features encompass risk maps, tables showcasing risk levels, graphs, charts indicating risk status, and a slider for area selection and querying. The interactive dashboard exemplified in Figure 2 displays drought risk levels in Ethiopia. These levels are derived from a weighted average of NDVI, SPI, and Rainfall Anomaly, with higher values denoting severe drought and lower values indicating minimal or no drought. Moreover, the dashboard offers insights into drought risk at the Ethiopian Woredas level. This visualization tool enhances accessibility and empowers users to make well-informed decisions regarding emergency management.

Figure 2: Dashboard for Drought Risk Mapping



5. Evaluation

The evaluation was conducted to assess the effectiveness and usability of the prototype, and to identify areas for improvement.

A total of 10 GIS experts completed the survey. All ten experts were chosen based on their experience and expertise in satellite image analysis, disaster risk mapping, PostgreSQL, PostGIS, and cloud computing. The findings are presented on the table below.

Table 1: Evaluation Criteria and feedback from the respondant

1	Data handling storage efficiency	90% of agreed or strongly agreed the prototype handle and store large volume of spatial data.
2	scalability	100% agreed or strongly agreed prototype can scle up or scale down to accommodate the growing volume of data
3	Security and privacy	60% disagree or strongly disagree the prototype have robust security feature and need to implement database role.
4	Interoperability	100% of respondents agreed or strongly agreed that the prototype can able to integrate with other system or software tools
5	Performance	100%) agreed or strongly agreed that the prototype have fast response times and minimal downtime.
6	User interface and usability	60% of the respondents agree or strongly agree that the prototype is user friendly but 40% disagree because rather than using using SQL script it is better to develop QGIS plugin. Also suggest to use machine learning model
7	Accuracy and reliability	80 % of respondents agree or strongly agree that the prototype able to provide accurate and reliable information. And 20% neutral because all the data are acquired from third party source and they agree that it is better to use data from real time sensors.
8	Cost effectiveness	100% of the respondents agreed or strongly agreed that the prototype is cost-effective, both in terms of development and maintenance costs

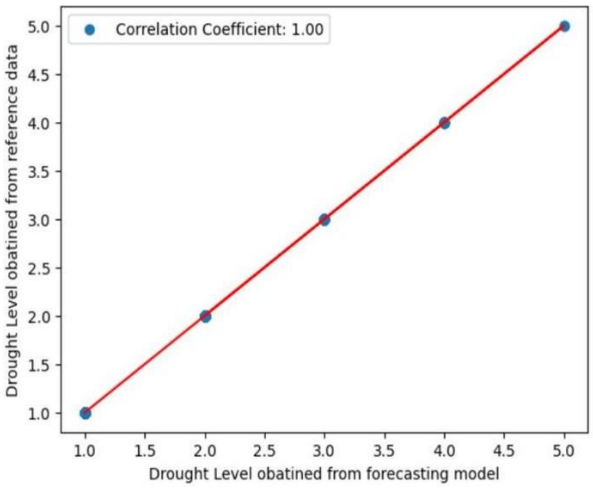
To assess the accuracy of a drought forecasting analysis evaluation metrics like correlation coefficient is used.

The correlation coefficient measures the strength and direction of the linear relationship between two variables (in this case, the drought forecasts and the reference data). It ranges from -1 to 1, where -1 indicates a perfect negative linear relationship, 1 indicates a perfect positive linear relationship, and 0 indicates no linear relationship.

The reference data for this research, to verify the accuracy of forecasted drought result, international databases like https://data.humdata.org is used. This websites have humanitarian data across the word and owned by UN OCHA including drought indexing of Ethiopia for year 2022.

By calculating correlation coefficient of the forecasted drought result of the research framework and reference data from the website sated above, using popular python on Jupiter notebook, the correlation result is 1 for all woredas in Ethiopia.

Figure 3: Correlation of drought by forecasting model and reference data using correlation coefficient



6. Conclusion and Future work

In conclusion, this research highlights the pivotal role of cloud computing in reshaping how we manage emergency situations. Cloud computing offers a secure, adaptable, and cost-efficient platform for responding quickly to unexpected and potentially dangerous events. The main goal of this study was to create a practical framework for handling emergencies using modern technology.

Through an extensive review of relevant literature, we've developed a four-layer framework that encompasses data gathering, storage, processing, and presentation. Each layer corresponds to a specific phase of emergency management: data collection, data storage, data analysis, and data visualization.

Our prototype testing validates the effectiveness of this spatial cloud data management framework. It efficiently handles large volumes of spatial data, scales to accommodate increasing data needs, provides easy access from different locations and devices, integrates smoothly with other systems, and does so in a cost-effective manner. However, there's room for improvement, including enhancing security and privacy measures, making the user interface more intuitive using a dedicated QGIS plugin, and exploring machine learning models to enhance accuracy in predicting droughts. Also, investigating more dependable sources of real-time data is a promising avenue for future exploration.

In essence, this research underscores how cloud computing can transform emergency management while offering a practical framework that aligns different data-related tasks. While the prototype performs well, ongoing refinements and advanced features are crucial to meeting the evolving demands of emergency response. This ensures a more secure, efficient, and robust system for safeguarding lives and resources during emergencies.

Reference

[1] Aly, A. G., El-Sharkawy, M. H., El-Bahnasawy, E. H., & El-Hadad, S. A. (2013). Proposed model of GIS-based cloud computing architecture for emergency system. International Journal of Computer Science and Mobile Applications, 1(1), 17-28

[2]Yang,C., Goodchild,M., Huang,Q., et al., 2011.Spatial cloud computing: how can the geospatial sciences use and help shape cloud computing. International Journal of Digital Earth, 4(4), pp. 305-329

[3] X. Ge and H. Wang, "Cloud-based Service for Big Spatial Data Technology in Emergence Management," in Proceedings of the 2011 IEEE International Conference on Computer Science and

Automation Engineering, Shanghai, China, 2011, pp. 324-328.

[4] W.W. Song, B.X. Sin, S.H. Li, et al., "Building Spatio-temporal Cloud Platform for Supporting GIS Application," in Proceedings of the 2015 IEEE International Conference on Big Data (Big Data), 2015, pp. 3112-3117. DOI: 10.1109/BigData.2015.7364094.

[5] D. Jiang and C. Yang, "Spatial Big Data Analytics for Emergency Management," in IEEE Access, vol. 3, pp. 890-898, 2015, doi: 10.1109/ACCESS.2015.2451843.

[6] S. Gao, X. Yuan and C. Zhang, "Spatial Data Management in the Cloud," in IEEE Transactions on Services Computing, vol. 6, no. 3, pp. 373-387, July-Sept. 2013, doi: 10.1109/TSC.2012.7.

[7] Demisse, G. B., & Rowland, J. D. (2014). "A spatial drought index for the Horn of Africa." IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 7(8), 3374-3384.

[8] Demisse, G. B., & Soliman, A. S. (2017). "Spatial analysis of drought in Ethiopia using the standardized precipitation index." IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 10(3), 966-972.

Publication 500-292." Available: https://nvlpubs.nist.gov/nistpubs/Legacy/SP/nistspecialpublication500-292.pdf

[9] Huete, A.R., Didan, K., Miura, T., Rodriguez, E.P., Gao, X. and Ferreira, L.G. (2002). "Overview of the radiometric and biophysical performance of the MODIS vegetation indices." IEEE Transactions on Geoscience and Remote Sensing, 41(6), pp. 1260-1266.

[10]Nigussie, A., & Tsunekawa, A. (2015). "Drought risk mapping using remote sensing and GIS: A case study of the Upper Awash River Basin, Ethiopia." IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 8(6), 2925-2938

Comparative Study of Smart Traffic Offence Ticketing System – Case of Addis Ababa City Traffic Management Agency

Tsegaye Atsbaha Gebretsahma
HiLCoE Graduate Programee, Ethiopia
itag143@gmail.com

Abstract

Recent statistical data shows that Addis Ababa City has registered more than 630 thousand vehicles. With that figure in mind it is possible to imagine the amount of daily traffic that the city could accommodate. It is becoming so common to see many drivers being stopped on the road and penalized by traffic officers every day. Since the current traffic offense management systems is almost completely manual, it is tedious and time consuming for a driver to settle an offence ticket and claim their license (whether it is vehicle's registration plate or driver's license) back. In Ethiopia, particularly Addis Ababa City police stations, there is no a comprehensive automated system that handles the document works related to traffic violation ticket system yet. Whenever a driver is caught up by violating the traffic road rules, they have to wait on the roads until the traffic officers write down fine papers manually, and then traffic officers ask to handover the drivers' license and/or vehicle registration plate. In order to claim their license, the drivers need to settle the fine in a bank or some agent, and take the receipt to the dedicated police station, and wait in a queue again. To intervene in this situation, this paper aims to conduct a comparative study of technological solutions for traffic offense ticketing system and recommend to Addis Ababa City Traffic Management Agency. Thus, the main focus is to compare four traffic violation ticketing systems: Traffic Ticketing Systems, E-Tickets, Automatic mobile ticketing system for traffic offences (Mobile Apps), and e-Traffic prosecution system. Based on the findings of the study, the researcher recommends that a customized version of traffic ticketing system and automatic mobile ticketing with integrated payment options best suits for Addis Ababa City Traffic Management Agency.

Keywords: traffic violation ticketing (TVT), traffic offense, e-payment

1. Introduction

According to statistical data, Addis Ababa city had around 150 thousand registered vehicles in 2014. During the period of this research, the city has more than 630 thousand registered vehicles. The increasing number of registered vehicles in Addis Ababa City is creating huge daily traffic flow on the roads. No wonder why the number of road traffic violations is increasing tremendously and so does the tedious manual process of traffic violation ticketing system.

An officer has to issue a manual ticket to the violating driver, and confiscates their license (vehicles registration plate or driving license) temporarily. The driver is supposed to pay the fine in a bank and/or agent, take the receipt as a proof to the concerned police station and claim their license back. This is so tedious, ineffective and insufficient process. It not only takes the time of the violating bodies but also waste a precious time and cumulative resource of all the participating entities.

The current system of traffic ticketing, which is almost completely manual ticketing system (of course,

there is some initiation to settle traffic offence fine though a mobile app, a comprehensive system is not implemented yet), is often bring inconveniences not only just to the deemed 'violators' but also one way or another all the traffic enforcers. To assist the police and the people with regard to improving the settlement of traffic tickets, a web-based or automated traffic ticket data information system is required, which can generate traffic ticket information quickly, at any time around the clock. This information system is aimed at reducing traffic violation committed by the people as well as increasing public awareness of safe, orderly and proper driving [2].

The fact that Addis Ababa City is accommodating large number of traffic violations day-in-day-out, there is a need to systematic implementation of traffic violation ticketing system not far from now. In addition, the paper dictates that a sustainable implementation of such systems may contribute a better decision making process based on accumulated traffic data on the long run. Such system implementation will provide an efficient, cost (time and resource) effective and reliable means of traffic offence data management process.

The present system of traffic ticketing, through the manual process, most of the time is totally inconvenient to both the traffic enforcers, the deemed violators, and public at large [6]. In Addis Ababa city, regardless of the ever growing road traffic violations, the TVT still falls under the old manual process. Addis Ababa City's Traffic Management Agency is also looking for technological solutions that could ease and automate the existing manual TVT systems [4]. The level of enforcement ranges from fine (in money) to driver license suspension and revocation depending on the severity of the traffic offence [1]. The purpose of the paper was to improve efficiency and effectiveness of traffic violation ticketing, to assist the police in addressing and reducing traffic offenses and at the same time speed up the way traffic ticketing process.

Therefore, on the course of this, a critical comparative analysis is undertaken in order to implement an effective traffic violation ticketing system through comparison analysis of the selected TVTs based on some criteria, and propose to Addis Ababa City. Hence, the objective of this paper is to compare selected traffic violation ticketing systems and recommend suitable technological solution that best fit for Addis Ababa City Traffic Management Agency (TMA).

2. Related Work

The use of technological solution has an impact on the settlement of traffic violation cases to replace the manual ticket system. With a smart traffic violation ticketing system, the drivers violating the traffic will be recorded through the application of the police. Increased of traffic violation is a huge challenge for the police to be able to implement educational sanctions while still having a deterrent effect [5].

A study conducted in Sri Lanka regarding traffic violation management system [9], points that the traditional paper based system is needed to be moderated and should be transformed using technological solutions. The new TVT systems overcome the drawbacks of the existing system by facilitating report generation as required to the auditing process in a very efficient manner. Moreover, with modern and traffic ticketing systems, future planning and decision making will be much easier. Based on the study, the researcher recommended the concerned authority should plan on modernizing traffic violation management system rather than wasting valuable resources on the current paper based manual process.

Traffic violation ticketing systems (TVT) is a broad concept that is used in wide range of systems for traffic management agencies and provide a centralized access for drivers, traffic police and city

administration. At high level, these services provide several types of traffic violation information such as type of violation, frequencies, area/location of violation and so forth. Beyond handing a violation ticket in an automated way, TVTs play significant role in delivering good traffic violation data management and can be used for further in policing and decision making processes.

Traffic violation management systems help to empower the penalty payment method regarding maintaining a reliable, swift and transparent payment method. As the system helps the traffic police to secure confidentiality of documents, report generation and view preview records of violations in a just a single window [9]. Starting small with just automating the traffic violation ticketing and fine settlement system, then adding more feature such as on-time GPS tracking of the police officer, push notification technology etc.…and gradually attain fully automated smart traffic management system could be attained.

According to the research, a deep observation and analysis is undertaken to better understand the possible problems and challenges faced while migrating from a completely manual process to an automated systems. Problems related to skill enhancement and training, awareness creation amongst stakeholders in regard to advantages of adopting technological solutions and resources constraints are majors ones. So, based on the findings, a summarized significance of the system for traffic violation ticketing as a technological management tool has already been deployed in many developed and developing countries and specially those cities that are trending on smart city technology

There are several technological solution alternatives when it comes to traffic violation ticketing system depending on the scope, budget, available resources (infrastructure and man power) as well as interests of all the participating organizations within. Researchers have conducted an intensive investigations on this field and have proposed variety of solutions. A research project conducted in Philippines, presents a solution on how to eliminate the long process of issuing traffic violations whereby a smart driver's license card will be implemented [10]. The research proposed an automated traffic violation apprehension payment systems in was that involves a cloud server with windows application, web based service and android application for users to interact with the system, and this was the effects of effective traffic management and apprehension of traffic violation endeavors.

An efficient and reliable traffic violation ticketing system is needed to create, settle and manage traffic violation tickets. TVT has a wide range of uses such as proper accurate reporting, data management and better decision making processes. A free and open source TVT system would be a good foundation to build a customized ticketing system. The system could provide a means to integrate an online payment options and city specific road traffic policies. Nonetheless, beyond adopting existing system 'as is', a systematic customization of selected system would be more preferable for a compatible and sustainable implementation.

Traffic ticketing systems is a fully automated web-based platform that facilitates traffic violation ticketing process. With its central database, this web-based traffic ticketing system is ideal for a swift and efficient traffic violation settlement process. The new traffic ticketing system provides an opportunity for the administrative settlement of traffic and road offences without having to go through the court system. Although this system takes an electronic approach, it does not totally eliminate the use of paper. Officers will be utilizing both the mobile handheld ticketing devices and the paper-based traffic ticket. There is now a newly designed traffic ticket that alleviates the current challenges experienced by law enforcement

officers when writing paper-based tickets. TTS, simply automates the common paper based process via a web based portal; that is after violation ticket is generated, the driver needs to settle the payment within the grace period of that ticket via bank or available payment option [1].

It is clear that bribery when traffic operations are frequent. That is the reason underlying the Police of the Republic of Indonesia implements a new system called E-ticket. A trusted system can reduce Pungli (extra money) and bribery practices. The E-ticket has been applied simultaneously to the simultaneous launching in Indonesia on December 6, 2017. The E-ticketing system will replace the manual ticket system using blank paper based ticket, where the offending driver will be recorded through the application [5]. Once and e-ticket is generated, the driver could settle the payment via bank, or using an online payment options. This Electronic traffic ticketing app, deals with handling the violation ticket process whereby police officer, issues the ticket and the driver receives the ticket on their phone.

With E-traffic mobile ticketing system, traffic enforcers should use handheld devices so as to issue a violation ticket. And the respective driver gets a ticket notification in real time via notification methods (on application, SMS and/or email). Since the police will use portable handheld devices to issue the violation ticket there is no an immediate need connectivity to central database. If internet connection is not available, the system generates and records the ticket on its local database and synchronize it when connectivity is restored. This ticketing system requires drivers to have the client side application installed on their phones and/or alternatively use digital account whereby they can access it in case of penalty settlement process. Automatic mobile ticketing system's main goal is to ease the ticketing process for everyone, regardless of where they are located. However, with the implementation of new technological solution challenges related to users' behavior to adopt new technology and system outages and/or availability issues are expected and should be planned ahead in time.

A flavor of automated traffic violation ticketing system, e-Traffic prosecution system, was launched and been successfully operated in Bangladesh since 2012. Main objective of the system is to bring transparency to the process of collecting traffic fines from vehicle owners and passers-by breaking traffic rules. Under the system, users will be able to check the documentation of their cars, driving license and other such documents through text messages. The traffic police can file cases against offender and the lawbreakers can also pay fine though a Robi (second largest mobile operator in Bangladesh)

With the availability of optional open-source traffic ticketing Systems, it is possible to customize one based on the business goals and agency's directives, allowing to get a best fit towards the needed purpose. Based on best practices and experts recommendations, let us have a look at the selected TVT's pros and cons in Table 1.

Table 1: Pros and Cons of selected TVT Systems

Selected TVT	Pros	Cons
Traffic Ticketing System, TTS	• Easy to use and set up via web • Migration to electronic ticketing from paper-based ticketing • Police need not to confiscate license and/or vehicle plate • Web based access	• Requires Internet Access for synchronization • Parallel paper work • Payment options need to be integrated

	• Track progress, get reports, share resources, review tickets,	
Electronic Traffic Ticketing (E-Tickets)	• Easy access to setup E-ticket application • Virtual license confiscation • Monitoring and report generation • Supports desktop and mobile devices • Avoids the use of fraud licenses.	• Needs a smart driver's license card in play • Payment options need to be integrated
Automatic Mobile Ticketing System	• Easy to use • Allows for multiple channel access partially • Customizable • Mobile apps available • Custom user interface • Simple profile management and data access	• Drivers' license should be digitally readable via handheld device (RFID, QR Code etc…) • Payment options should to be integrated
E-Traffic Prosecution System	• Simple process for police • Law enforcement officer's labor work minimized to almost 98.9% • Good for accessing detail documentation and user data • Web based access for both users and law enforcement bodies	• Needs an agent/call center for payment handling • Payment option should be integrated

The advancement of information and communication technology is affecting many aspects of our lives. The internet and the World Wide Web innovations are increasingly promising to provide various tools to increase socio-economic developmental process for developing countries like Ethiopia. Addis Ababa being the capital of Ethiopia, is highly populated city and with more than 600 thousand registered vehicles. It is presumably fair to say that there expected increase to the number of traffic violation occurrence rates.

Traffic violation has become an ever growing event and with the frequency of violations, there is a demand perhaps forced one, to change the way ticketing of traffic offences are handled. Improving speed, information management, reducing operational costs and human errors, and enhancing accountability of receipted money. Automatic ticketing system is the automation of the manual ticketing process currently used by the police traffic officers [7].

Based on the current traffic flow and observation made on the means traffic ticketing system in Addis Ababa, there definitely is a need to intervene with currently existing technological solutions, but of course with some adjustments to comply with the city's traffic management policy. The information technology development in Ethiopia, together with the fast growing mobile adoption and usage reaching over 56 million achieving more than the subscribers based target, brings an easy way to adopt technology based applicable solution to the traffic ticketing system. However, adopting a new technology is not primarily a technical problem but rather a social problem, where customers' acceptance plays an important role [8]. Thus, adopting and/or developing a solution that provides accurate, timely and efficient process of traffic

violation ticketing is a need, rather than just a trend to glare at.

Although literature is rich with e-business solution, e-traffic solutions have been overlooked [8]. The spread of traffic violations are accompanied in service provision, control monitoring, fine ticketing, changes in driver behavior, and changing relations between citizens and the police officers. Generally, cities with high population and great number of vehicles, like Addis Ababa, should normally have a greater capability to increase the economic productivity of the society.

Smart traffic violation ticketing, TVT, is the process whereby users (the traffic officers, the violating drivers and/or the traffic management office) depending on their role can order, pay for, obtain and validate tickets from any location and at any time using mobile phones or other handheld devices. TVT, could be defined as a method by which law enforcement agencies use in-car computers to generate traffic citations in the field, and the print an electronic or a hard copy for the offender [8]. This way a better and swift way of traffic ticket and later report generating means would be attained. Later on data generation, aggregation, management and decision making process would more accurate and help on making crucial traffic policies.

An investigative study focused on mobile traffic violation ticketing services in Egypt [8] reviewed several possible technological solutions available and intuitive to modernize traffic violation ticketing system using analytical tools. It persuades that a significant relation has been found between the new service ease of use, perceived usefulness, compatibility, mobility, and drivers' perspectives with regards to their intension to adopt the mobile traffic violation ticketing services. In Egypt, with more than 100 million population and large number of registered vehicles, the interviews believe that the ministry's perspective shows adopting TVT service will help to improve the ticketing system immensely. It is a way of developing all traffic departments, helps drivers an immediate notice of violation and an immediate deterrent, may reduce irregularities and facilitate traffic violation ticketing service for drivers. However, there is no such smart traffic violation ticketing system for nationwide implementation, yet.

Rino et al [12] studied factors affecting users' acceptance of e-billing states that the level of acceptance of the user to use the new system determines the success or failure of these systems. As a result of their findings two important factors; performance expectancy and social influence affects system usage significantly. Therefore, understanding the factors that affect system acceptance can help the government/administration to formulate appropriate measures to promote better e-billing services to the public. This won't be much different when it comes to adopting new technological solution for traffic violation ticketing system.

By the comprehensive report generation module authorized personnel can get detailed information about the traffic violation such as district, location, area wise etc. In general such TVT system, as expected, facilitates all the requirements from each and every stakeholders' perspective. With an integrated secure payment method through the system to the redeemed violators, reduces his/her responsibility and facilitates the payment process to be quick. Almost all of these system database follows well known standards to secure user information that brings confidentiality of relevant user details. Proposed system mainly improves accessibility of users specially police officers and guilty parties to make ease of their process. The proposed system has a great advantage in report generation as required and overall auditing process [9]. Forecasting process, auditing and future planning are the commonly expected added features of such a system.

3. Comparative Analysis and Recommendation

For many cities, the successful deployment of traffic violation ticketing system has been the foundational element on the course of overall traffic flow management. For developing cities like Addis Ababa, this is not that case, rather opt to deal with the old resource intensive manual process. At some point there is a need for change so as to catch up with ever growing demand of all the involved parties be it the deemed traffic violators, enforcing officer or supplementary traffic related offices. When it comes to choosing a suitable TVT systems, there are several factors that should be considered such as infrastructure, work flow automation, responsiveness & speed, availability, interface/usability, and importance of the system/collaborative.

While the usage of these systems gains recognition and acceptance in many cities, there are new problems arising that need to be solved. Because of multiplicity of platforms and approaches for systems implementation, it becomes increasingly difficult to manage or compare them. Each new TVT presents its own learning curves.

Comparing different traffic violation ticketing systems, and on what criteria to opt the most suitable one, is a task of ever-increasing importance. With the E-ticket, it is easier for the public to pay a fine through the bank. However, not all communities can follow the E-ticket procedures provided by the agency, especially those people that deter from using such technology [5]. This paper studies and compares some widely known traffic violation ticketing platforms: Traffic Ticketing System (TTS), E-Traffic, e-Traffic Prosecution System and Automatic Mobile Ticketing System. The final results show that the operation of the traffic ticketing system in reducing the cost of ticketing, improving information management, improving the speed of ticketing process and reducing in human errors potential be improved.

Whenever there is a need to an automated solution, the fundamental task is to decide on the kind of deployment whether it should be cloud-based or on premise /standalone ones. Essentially, the fundamental difference between the two is where it resides. On-premise software is installed locally, on your business' computers and servers, where cloud software is hosted on the vendor's server and accessed via a web browser. Cloud service has several advantages like reducing IT staff’s responsibilities, adjust to your budget, regular data backups, adjustable to needs, and eliminates capital expenses. However, cloud based service has disadvantages cause data is less secure, access is based on connection (fast and reliable internet is required), cost can balloon with little warning, and even litigation.

Unlike cloud based service, on premise service relies on own infrastructure, meaning you have to own all of the equipment, and be responsible for the lifecycle and operational management. On premise storage is preferable the fact that it can operate without internet, lower monthly cost, greater security, control over server hardware, data confidentiality and so forth. Nonetheless, it requires extra IT support, adherence to industry compliances, increase maintenance costs, risk of data loss, less scalability and requires greater capital investment. Thus, the decision remains to be which way to go depending on the implements requirement and IT policies of the concerned organizations and authority.

When adopting new technology for traffic violation ticketing and management system, it is critical to compare each and every feature and available services even though there may not even a comprehensive framework to do so. With the best available TVT systems availability, an organization can adopt any of the mentioned alternatives as is and/or with some customization, to best fit their traffic offense ticketing

process.

Based on the needs and policy based requirements of organizations (scope of implementation) and the fact that most TVTs share common core features, it would be challenging to choose one over another. Nonetheless, Table 2 shows a tabular compiled illustrations of fundamental features and components that should be present in a decent traffic violation ticketing system that are selected in this study.

Table 2: Comparison of selected TVT systems

	Traffic Violation Ticketing	Electronic Traffic Ticketing	Automatic Mobile ticketing	E-Traffic Prosecution System
Collaborative	Yes	Yes	Yes	Yes
Responsive and Quick	Yes	Yes	Yes	Yes
Self-help	Yes	No	No	No
Search friendly	Yes	Yes	Yes	Yes
Reports & Analysis	Yes	Yes	Yes	Yes
Workflow Automation	Yes	Yes	Yes	Yes
Multilingual Support	Yes	No	Yes	No
Custom user interface	Yes	Yes	Yes	Yes
Payment Integrated	No	No	No	No

According to [8] the traffic violation management system demand is a hundreds of million dollars project when applied in full scale. Developed nations have already deployed smart traffic management systems, fully automated and integrated systems. In this paper, however, the scope limited to traffic violation ticketing system (TVT), and when you choose the right TVT for your city or country, you ease the overall traffic data management which in return facilitates in data reporting, monitoring, policing and decision making process.

There are plenty of ways that a traffic violation ticketing system could be selected and deployed for action. But just before selecting such a system, it is crucial to choose the best suitable deployment for that specific TVT to be preferred. Technological experts argue that self-hosted and/or standalone systems are becoming less popular over their cloud based counterparts. Yet, self-hosted systems have their own advantages as compared to cloud-based ones. Thus, type of deployment is essential for factor when comparing the available systems as it impacts the overall working process one way or another.

Considering the current internet connectivity coverage and accessibility situation in Addis Ababa, it would be wise to use on premise deployment type over cloud services. Even though on premise systems demand higher infrastructure and running cost initially, it would more advantages for a reliable and available systems to be operated. When it comes selecting a suitable TVT system, for several reasons mentioned in previous sections, features that make life easier for both the administration and driver could be compromised in consideration of the city's traffic management policy as well as digital payment regulations. Most TVT share similar features but differ in their way of logical implementation. Some are web-based, some are APIs and mobile apps, and some are software as a service platforms. Traffic violation ticketing system, mentioned above, is so intuitive and easily customizable for web based access. The other

technologies orchestrate their own flavor.

When looking at the type of system deployment, from the researcher's point of view, on premises hosting may bring frequent downtimes of servers due to the cities frequent power outages, service interruption and related factors as a challenge to a successful TVT deployment. In the case of cloud-based system implementation, higher cost for hosting and technical support and information confidentiality might be the challenges to be concerned about. Thus, there will be a need to careful planning and compromising on how to deploy the system when adopting it in the first place.

Based on [11] with the emergence of e-governance and e-service, many cities has adopted traffic violation ticketing system, even able to attained fully automated traffic management system that integrated with their smart city. When it comes to new adoption of a technological solution, users tend to resist e-service if the providers do not ensure the accuracy and reliability. Users will be reluctant to adopt e-governance systems if they do not trust governments and their provided services. Since traffic violation ticketing includes sensitive information such as transactions and involves the disclosure of personal information, trust is believed to be a plausible determinant towards a long term use. Moreover, users do not want to invest on new devices and/or resources whenever there is a new technology to be used.

Based on this comparative study approach, a TVT system that is based on mobile app with web service support and integrates various payment options is presented. Integrated database driven and web based access features from traffic ticketing system and standalone mobile application depending on the OS (whether for android, iOS or others) would bring more convenient way of system implementation. In addition to that, payment means needs to be integrated to this system so that the drivers could easily and access their traffic offices and overall transactions from a single window. Thus a customized version of Traffic Ticketing System, TTS and Automatic Mobile Ticketing System, i.e. a hybrid solution, is suitable for traffic violation ticket system for Addis Ababa City. Making the system accessible via multiple channels like web-based, smartphone apps and so forth will be an advantage for the users (drivers) to get them easily fall into the system.

To summarize all the above comparisons, a custom TVT system that is a hybrid of traffic violation ticketing and automatic mobile ticketing system with an integrated online payment options, is suitable for Addis Ababa City Traffic Management Agency.

4. Conclusion

With the ever growing need of humans for mobility and use of vehicles, traffic management authorities are using various traffic violation ticketing systems to issue and manage violation tickets in a swift and efficient way. Most countries in the world have deployed different advanced traffic violation ticketing systems effectively. In Addis Ababa, Ethiopia, regardless of the ever growing road traffic violations, the TVT still falls under the old manual process. Addis Ababa City's Traffic Management Agency is also looking for technological solutions that could ease and automate the existing manual TVT systems. Observing the above challenges, the research conducted comparative study and recommended TVT system that best suits to the city's traffic violation ticketing process. As a result of the possible technological solutions amongst of the selected tools: a hybrid solution of Traffic Ticketing System and Automatic Mobile Ticketing System with an online payment options integrated, is recommended.

References

[1] Gadisa Layo Mosisa, "Analysis of Road Traffic Violations in Addis Ababa City (The cause of Arada sub-city)" AAU/AAIT/ MSc Thesis, 2018 G.C

[2] Faikul Umam, Rachmad Hidayat, "Web-Based Traffic Ticket Data Processing Information System", BISSTECH 2015

[3] Widad M Faisal, Shehab Al-Sakkaf, Aymam Baalwi, …Abobakar Alattas, "An Automatic Traffic Violation Capturing System Using Radio Frequency Identification(RFID) Technology in Mukalla City", Oct 2018

[4] "Ethiopia: Agency Debuts Traffic Fines Online Payment." [Online]. Available: https://allafrica.com/stories/202104210506.html. [Accessed 25-June-2021].

[5] Sri Endah Wahyuningsih, Muchamad Iksan, "The Benefits of the E-Traffic Ticketing (E-Tilang) System in the Settlement of Traffic Violation in Indonesia", 2019

[6] Angilyn Leoncio, Florocito Camata, "MMDA E-Ticketing System", Dec 2017

[7] Gershom Chishala, Dr. Alice P Sgemi, Maybin Lengwe, "Automatic mobile ticketing system for traffic offense", July 2018

[8] Amer, H., Abd El Aziz, R. and Hamaz, M, "Investigating mobile traffic violation ticketing service in Egypt: The provider and user perspectives", June 2014

[9] Sandas Tharinda, "Implementation of online motor traffic violation management system", 2020

[10] Luisito L. Lacatan, Paul S.Gilber, Seth B. Guerrero, Ken L. Pascal. Patrick M. Tolentino, "A Device for Integrated Traffic Violation Apprehension and Payment System Using NFC Technology", 2017

[11] Omar E M, Khalil and Ala'a Al—Nasrallah, "The adoption of the traffic violation E-payment System (TVEPS) of Kuwait", 2014

[12] Rino Ardhian Nugroho, Arlyn Dewi, Arum Pratiwi, "Factors Affecting Users' Acceptance of E-Billing System in Surakarta Tax Office", May 2018

[13] Sentry, "Self-Hosted or Cloud Sentry?" 2021. [Online]. https://sentry.io/resources/self-hosted-vs-cloud/. [Accessed: 25- Sep- 2021].

[14] Rasha Abd El.Aziz, Rehaballah, Miran, "ATM, Internet banking and mobile banking services in a digital environment: The Egyptian banking industry", 2014

[15] Staff Correspondent, "e-Traffic prosecution system launched in capital" [Online]. Available: https://www.thedailystar.net/new-details-251134. [Accessed: 02- Oct - 2021]

Distributed Databases for Better Performance, Scalability, Resilience and Transactional Integrity in the case of MOENCO

Hiwot Workneh Denekew
HiLCoE Graduate Programee, Ethiopia
hiworkneh@gmail.com

Abstract

There are several factors to consider when choosing the right database engine for an enterprise, and that depends to a large extent on the type of business the data base is intended to serve. In the case of MOENCO S. Company, which is the main focus of this research, a number of factors are considered, including cost (storage and writing charges in particular), where they value consistency over size, making it incredibly expensive, and (speed over consistency which results in data discrepancy). When selecting a database management system, there will always be tradeoffs to consider, and one size does not fit all. However, there are new database products released that take a lot of inspiration from the NoSQL focus on performance, distribution and scalability and without throwing out all the relational features that have been useful for so many years. These products which include Google's Cloud Spanner and Apache Ignite are sometimes referred to as NewSQL, replacing earlier products such as NoSQL that support ACID transactions instead of SQL queries, and generally following the relational database model while maintaining features oriented towards scaling and speed, much more like the NoSQL products.

The purpose of this paper is to conduct a comparative analysis of cutting-edge database technologies that adopt a multi-model approach and based on the analysis of findings recommend a more viable database solution that is an open source, high performance, distributed database engine, and could serve better the achievement of the organization's objectives while addressing the challenges at hand. This paper also addresses various challenges and benefits in implementing relational and non-relational (NoSQL) databases as well as address the designing factors in Distributed SQL databases like scalability, resilience, and transactional integrity across selected database engines and review of recent work.

Keywords: Distributed SQL, RDBMS, Resilience, Scalability, Geo-distribution, NewSQL

1. Introduction

A database is a logically organized collection of structured data kept electronically in a computer system. A database management system is usually in charge of a database (DBMS). The data, the DBMS, and the applications that go with them are all referred to as a database system, which is commonly simplified to just database. A database engine (sometimes known as a storage engine) is the fundamental software component that allows a database management system (DBMS) to generate, read, update, and delete (CRUD) data from a database. The terms "database engine" and "database server" or "database management system" are commonly interchanged [9].

Databases are a concept that we are all familiar with, and we have all dealt with them; they are difficult to escape if we work or do anything with computing. Databases allow us to operate efficiently while working with enormous amounts of data They make data updates simple and dependable, assist in ensuring accuracy and avoiding redundancy, and provide security measures to manage access to information. There are different kinds of databases engines or database management systems that were intended for different use cases and choosing the right database and the right storage platform to fit the

business needs can shape the success of the entire organization [10].

A database is commonly described as a repository of data or a place to store your information, which is technically correct, but it is not why it is valuable, it is not just because we have data that we need databases, it is for what comes after that and there are major issues and requirements that need to be taken into consideration in this regard such as [11, 12]: what will happen when the data grows, how keeps track of the data when it changes over time, how can our data be shared, fast and made searchable, and accessible, how will the data be reliable, available and credible (Data integrity), how can data be consistent with itself, how can data be protected and secure at all times?

Keeping these requirements in mind, it is critical for every organization to select an appropriate database engine to assist them in dealing not only with a large volume data they need to process and manage but also with all the additional issues that having a well-established and functioning database management system constantly entails. It will be up to the company to determine what is most important the organization considering the requirements discussed above [11].

Databases are one of the few technology products released 30 - 40 years ago that are still relevant today. For instance, IBMs DB2 which was released in 1983 has remained a viable product and a relevant skill ever since. In addition, PostgreSQL, which is a popular open-source DBMS, was first released in the 1980s as was Microsoft SQL Server and Oracle that was released in 1978. Even though these products have been improved, updated, and refined over the years they have not lost their continued relevance for decades. These products and other popular DBMSs were built on the same set of ideas and concepts, and they share a similar approach to what a database should be and belong to a particular type of database management system called relational database management systems (RDBMS). While there are other DBMS types, relational ones have been the most popular and most prevalent databases over the last several decades.

The reason why the relational database management systems (RDBMS) products have stayed relevant for decades is because they are very good at what they do and what they do is perfectly suited for many typical business scenarios. But in the last few years, we have seen various amounts of new products released in this space and many more recent database technologies that were designed to deal with specific situations these standard relational databases were not great at. In order to explore and appreciate the added value of the newer ones it is necessary to compare and contrast them with classic relational databases. For that reason, some of the features of the classic relational databases are addressed in this paper.

The research report will begin by describing the characteristics of classic relational database management systems and why they have remained so popular over time, as well as their limitations. Then it compares and contrasts the current marketplace database solutions including other leading database engines present at the moment. Finally, the research report will recommend a cutting-edge database engine with features drawn from both relational and non-relational approaches.

Hence, the objective of this research paper is to conduct a comparative analysis of cutting-edge database technologies that adopt a multi-modal approach and recommend a better database engine that is suited for improved performance, scalability, flexibility, integrity and to propose a solution that is appropriate for the problem domain.

2. Related Work

Database management systems (DBMS) fall into various broad categories. The most common and widely used DBMS is what is called a relational database management system (RDBMS). Other types of database management systems include Hierarchical DBMS, Network DBMS, Object-Oriented DBMS, including a more recent ones that fall under the category of NoSQL Database systems [6].

The relational database management system has been such a consistent part of computing because the features it provides such as, having strict controlled schemas and ACID transactions [6]. These features while working in a lot of general-purpose business scenarios, may not be ideal for every situation. The same database feature that can be hugely beneficial in one scenario can be detrimental in another. In today's world of Agile development where new features and updates are being released and implemented in cycles of weeks or even days, making continual changes to a database schema can become a challenging task. Beyond that, a couple of decades ago, there were no big data sets and real time web applications and there were no expectations for constant availability and global deployment (i.e., taking a database and replicating it across multiple international data centers) [4]. Over the last several years however, additional database products immerged that were intentionally created to deal with these situations relational databases were not designed to deal with, such as, flexible schemas, massive scale and geographic distribution [1].

A distributed SQL database has three important features. these are, Resilience, Scalability and Geographic distribution [2]. While resilience means failure should not affect the user or there should be an ability to return to a previous state after the occurrence of some event or action which may have changed that state, scalability means adding more nodes in order to do more aggregate processing, i.e., weather it is for a number of queries one is able to handle or total storage volume. Geographic distribution, on the other hand, refers to multizone deployment (i.e., replication of data across multiple servers on different locations) where one is able to service zone failures with low latency.

Distributed SQL has the following five defining characteristics [1,4,6],

1. *It should have an SQL API*: This means developers or users of DBMS such as MySql, PostgreSQL or an Oracle should be able to use their knowledge of SQL development and directly apply it to distributed SQL systems. In such cases, there is no need for learning a completely brand-new language in order to get data in or out of a distributed SQL system.
2. *Distributed Data Storage*: This is a case where data can be inserted easily into the system and depending on the replication factor and the number of nods in the cluster, the system should be intelligent enough to know how to distribute that data so that there will not be any human interaction or involvement.
3. *Strongly Consistent Replication*: This means that whenever data is inserted into the system, data integrity is maintained and it's the data that's not eventually replicated to the other nods where the it needs to reside. This is done in a synchronic and consistent fashion were, regardless of communicating with node 1, node 2 or node 3, there will be a consistent view of data in the system.
4. *Distributed Query Execution*: This means there is no need to figure out where the data is and how to pull it together and get it over to the client among a distributed data across a multitude of nodes in a cluster. Instead, one can easily run a Select, an Insert and an Update query and have the system do the rest of the heavy lifting.
5. *Distributed ACID Transactions* [7]: ACID transaction is a well-known concept in traditional RDBMS

but the only degree of complexity that is added here is that this ACIS transactions have to be aware that one is communicating over different nodes, dealing with data that may or may not be present on specific nodes and being able to roll back these transactions. All these aspects need to be taken into account when dealing with ACID transactions inside a distributed system. Hence, support for distributed ACID transactions means, a distributed SQL database automatically distributes and strongly replicates data so that SQL can be executed against it in a distributed and ACID compliant manner.

With the Internet and cloud computing gaining popularity, new types of applications have evolved, increasing the demand for database technology, like low latency and a high concurrent reading and writing rate, Big data storage and access requirements that are efficient, high scalability and high availability, lower management and operational costs, slow reading and writing, limited capacity as existing relational database cannot support big data in search engine, Big System and Multi-table correlation mechanism which exists in relational database.

A variety of different types of databases have emerged to address the aforementioned needs. Because all these novel databases differ significantly from typical relational databases, they are known as "NoSQL" databases. NoSQL can also be translated as "NOT ONLY SQL" to emphasize the benefit of NoSQL [13].

Any non-relational database is referred to as a NoSQL database, and there are many distinct implementations of NoSQL, including table formats, documents, and graphs. The following are the main benefits of NoSQL: 1) fast data reading and writing; 2) mass storage capability; 3) easy expansion; 4) inexpensive cost

On the grey side, NoSQL has several shortcomings, including lack of support for SQL which is an industry standard, lack of transactions, reports, and other additional capabilities, and the fact that most NoSQL database systems launched in recent years are not mature enough. Following an overview of NoSQL database, below is a comparison of Distributed SQL with NoSQL based on the requirements stated before [14].

Table 1: Comparison of Distributed SQL and NoSQL [6, 14]

6.

Characteristic	**Distributed SQL**	**NoSQL**
Performance	Slightly slower read and/or write performance as it is relational	High performance read and/or write
Scalability	Are designed to scale up vertically for more datasets size.	Datastores are designed to scale out horizontally
Resilience to failure	Master-slave replication, which makes the strongest resiliency power	Peer-to-peer replication
Transactional Integrity	Have natural atomic transactions (based on using ACID)	Provide eventual consistency from BASE (Basically Available, Soft State, Eventual Consistency) properties
Structure	Structured tables to store the data	Semi/Unstructured data collections that are not related to each other

From the literature review made above it can be observed that there are emerging opportunities to link

the relational SQL database with NoSQL – Non Relational database features that would allow product designers to effectively address the growing needs of users in terms of balancing performance, scalability, resilience and transactional integrity.

MOENCO is a subsidiary company of Inchcape PLC, a London based company engaged in global distribution & retail leader in the premium and luxury automotive sectors dealing with more than 12 brands of automotive and machinery.

As a leading global automotive dealer, distributor and retailer in Ethiopia, the firm has a high volume of transactions originating from each of its ?? branches in the country and where all of the distributed branches need to be updated from the main branch (particularly in TOYOTA spare parts sales). Standard DBMS features and requirements such as performance, scalability, transactional integrity and resilience to failure are very important for the company's day to day business management. In addition, the business requires high synchronization (quick update and delete statements are required to ensure that data integrity is maintained) and high-speed to adequately respond to the needs of its large and ever-growing customer base.

MOENCO had been using a locally developed tailor-made ERP system called Hillmark which has served the main office and Its branches for over a decade with a distributed system architecture - traditional relational database (Microsoft SQL Server2008-2012). Through time however, the company began to experience significant storage expenses as well as speed concerns as the volume of its data grew. As of the recent March 2019 to be precise, the company switched to a K-ISAM flat-file non-relational database (ADP Autoline Database Filesystem) called CDK Autoline DMS, which is faster than the earlier versions but still presents several challenges. Among these challenges include, data redundancy, requirement of several entries, the fact that specific data is erased when the system archives, cases where transactions are lost owing to lack of data harmonization, and common occurrence of read/write conflicts, to name a few.

However, there are new database technology products released, similar to Google's Cloud Spanner and Apache Ignite that include both features (hybrid databases) and are therefore ideal database engine in the context of MOENCO's requirements. The company requires speed because it's sales (large automotive and after-sales transactions) are dependent on it. It also requires data consistency because the company must perform analytics, study customer data and behavior in order to increase sales, as well as provide data to the government and report to its parent company. This research paper aims to conduct a comparative analysis of cutting-edge database technologies that adopt a multi-model approach with a view to recommend improved and state-of-the art technology solutions that would allow the company to meet its business objectives and meet the ever growing demands of its customers.

3. Comparative Analysis and Recommendation

When it comes to selecting the ideal database for an enterprise, there are several factors to consider. These include different database options that could generally fall into broad data categories, such as Relational, Documented, Key-value, Graph, In-Memory, search, time series, and many more others [7]. The comparative analysis here will focus on the following key database technologies, where each of these database technologies are built for a specific purpose and functionality, making them somewhat unique and different in their own way.

MySQL is a multi-threaded, multi-user SQL (Structured Query Language) database server that is exceptionally dependable [6]. It is intended for use in mission-critical, high-volume production systems, as well as in widely dispersed applications. MySQL has a number of downsides and limitations, including

licensing and proprietary features. It is a dual-licensed software, which exists in a range of commercially available versions under proprietary licenses. As a result, several functionalities and plugins are only available in the paid editions.

Amazon Aurora has been generally available since 2015 as stated on and is based on a proprietary distributed storage engine that replicates 6 copies of data across 3 availability zones for high availability [1]. It is compatible with both PostgreSQL and MySQL in terms of API. Aurora runs in a single-master configuration by default, which means that only one node may process write requests and the rest are read-only replicas. Multi-master configuration is a new feature in Aurora MySQL (although not yet in Aurora PostgreSQL) that allows you to scale writes by adding a second writer node. However, because both nodes now have all of the data, concurrent writes to the same data on both nodes might cause write conflicts and deadlock problems, which the application must deal with. The inability to scale beyond the original two writer nodes, as well as the lack of geo-distributed writes across several regions, are just two of the many drawbacks [1]

PostgreSQL is a powerful object-relational database management system that includes transactions, foreign keys, subqueries, triggers, user-defined types, and functions. PostgreSQL creates a new process for each new client connection [1]. Each new process is given about 10MB of memory, which can quickly add up when dealing with large databases with many connections. As a result, PostgreSQL is usually slower than other RDBMSs, such as MySQL, for simple read-intensive operations [2], which makes low memory performance its major limitation.

Spanner is Google's worldwide distributed SQL database that runs several of the company's mission-critical services, including AdWords, Apps, Gmail, and others. Since 2007, Google has been developing Spanner, initially as a transactional key-value store and then as a SQL database. In 2017 [3], Google Cloud Spanner, a private managed service, made a part of the Spanner system publicly available on the Google Cloud Platform. Cloud Spanner has its own SQL API that isn't compatible with any other open-source SQL database and doesn't support relational data modeling components like foreign key constraints [3], which appears an advantage from Google’s perspective but can be considered as a limitation from users’ perspective.

The sharding, replication, and distributed transaction features of CockroachDB, which is a PostgreSQL-compatible distributed database, is inspired by Google Spanner. Because the API layer is a reimplementation of the PostgreSQL query layer, functional depth is lost. Partial indexes, stored procedures, and triggers, for example, are not supported. CockroachDB also recently abandoned its open-source roots by relicensing the primary database under the commercial BSL 1.0 license, which restricts freedom of usage. Finally, complex functionality such as distributed backups and change data capturing are not available in the premium edition [5] that could prove to be costly to users.

YugabyteDB was inspired by both relational SQL concepts with the Non-relational- NoSQL concepts by using state of the art technologies in both aspects (PostgreSQL and Google Spanner) by considering the fact that PostgreSQL is more popular and fast-growing DB than any other database, and Google spanner is a truly, horizontally scalable and strongly consistent relational database that is cloud native [2]. When looking at the two sides however, the PostgreSQL is not cloud native meaning it cannot do scale, high availability in spite of failures, geographic distribution, distributed transactions etc. Google spanner does not have all the RDBMS features set and the quality that PostgreSQL has but putting these two together is the best way to build the perfect distributed SQL database which is Yugabyte DB.

The main difference between monolithic, single node databases like MySQL or PostgreSQL and

distributed SQL databases like Yugabyte DB is that by default in a monolithic single node database, all writes need to be served from a single node. In order to get more scale from databases like MySQL or PostgreSQL, one has to add more memory more CPU and more storage capacity, in other words 'Scaling Up'. On the other hand, with distributed multi-node SQL databases like Yugabyte DB, writes can be served from any node and in order to get more scale from a distributed database one needs to add more nodes to the clusters, in other words 'Scaling Out' [2]. With that definition in mind Yugabyte DB is built to be a distributed SQL database with a focus on high performance which means the ability to do relatively low latency serving and ability to deal with large amount of CPU, memory, Ram etc. It is an OLTP database and its cloud native so it will run anywhere and doesn't have external dependencies and the database itself is open source just like PostgreSQL.

Table 2: Comparison of PostgreSQL and Google Spanner

Characteristic	**PostgreSQL**	**Google Spanner**	**Yugabyte DB**
Query Layer	Very well known	New SQL flavor	Reuses PostgreSQL
SQL Feature Depth	High (Many advanced features)	Low (Basic features missing)	High (Supports most Postgres features)
Scalability	No Horizontally scalable	Horizontally scalable	Horizontally scalable
Replication	Async	Sync, Reads replicas	Sync, Reads replicas

As shown in Table 2 above, when comparing the query layer and the SQL features we observed as a part of PostgreSQL are well understood, adopted, known and really advanced. On the other hand, Google spanner misses triggers or stored procedures or portal partial indexes a whole bunch of long list of features and has a very basic feature set and it's a completely new flavor of SQL. When we look at what Google Spanner has done though, it has high availability, horizontal scale and distributed transactions. It also does Synchronous replication while PostgreSQL and most of the RDBMS do Asynchronous replication. So with Yugabyte DB which it a completely open source database just like PostgreSQL, has a combination of the best of those two worlds (the best of PostgreSQL and Google Spanner) [2].

PostgreSQL is a single-node open-source RDBMS, that is widely adopted for its powerful set of features while being extensible [2]. However, it is difficult to run PostgreSQL as a cloud-native database to inherently survive failures, by highly available, scale horizontally, and be deployed in geo-distributed configurations. Google Spanner is a distributed SQL database that has these features, however, does not offer the power of PostgreSQL. Combining the best of these two databases would result in a very compelling database. YugabyteDB is a fully open-source distributed SQL database aimed at achieving exactly this goal [1].

Table 3: Comparison of Amazon Aurora, Google Spanner and Yugabyte DB

Feature	**Amazon Aurora**	**Google Spanner**	**Yugabyte DB**
SQL Wire Protocol	Yes (PostgreSQL, MySQL)	No (New SQL flavor)	Yes (Reuses PostgreSQL)
RDBMS Feature Support	High (Many advanced features)	Low (Many features missing)	High (Supports most Postgres features)

Fault Tolerance	Yes (Multi zone/ multi region cluster)	Yes	Yes (maximum number of node failures it can survive while continuing to preserve correctness of data.)
Scalability	Writing is not horizontally scalable (vertically extend all write nodes or master nodes to extend the write throughput)	Horizontally scalable	Horizontally scalable
Transaction concepts	ACID compliant	ACID compliant	Distributed ACID with Serializable & Snapshot Isolation. Inspired by Google Spanner architecture.
Replication	Async	Sync, Reads replicas	Sync, Reads replicas

As shown on Table 3 above, Amazon Aurora reuses PostgreSQL and MySQL as its query layer but Google spanner builds something new from scratch and the feature depth that Amazon Aurora has are much high and supports the RDBMS features than Google Spanner and Yugabyte DB has a combination of both.

From the perspective of cost and speed, which are the two main requirements for efficiency, effectiveness and profitability in MOENCO S.Co. the use of Yugabyte DB is recommended both wither to be deployed on premises or on any cloud platform. The fact that Yugabyte DB supports advanced RDBMS features like triggers, stored procedures, foreign keys and some Postgres extensions and all distributed transactions are guaranteed Atomicity, Consistency, Isolation and Durability, in other words ACID transactions [7], makes it even more attractive as a viable database product.

Table 4: Comparison of Distributed SQL Databases

Characteristics	**Amazon Aurora**	**Google Cloud Spanner**	**Cockroach DB**	**Yugabyte DB**
SQL & Transactions Compatibility	PostgreSQL and MySQL (Full Compatibility)	Proprietary SQL (No Foreign Keys)	PostgreSQL (No Partial indexes, Stored Procedures and Triggers)	Reuses PostgreSQL (Full Compatibility)
Native Failover/ Repair	Automated Repair	Automated Repair	Automated Repair	Automatic repair of missing data after failures, using unaffected replicas as sources
Horizontal Write Scalability	Yes (Multi-Master)	Yes (Auto-Sharded)	Yes (Auto-Sharded)	Yes (Auto-Sharded)
Geographic Data Distribution	Yes (Only a single region can take writes)	Yes (Global clock sync for consistency)	Yes (Global clock sync for consistency)	Yes (Global clock sync for consistency)
Cloud Neutral	No (Proprietary to AWS)	No (Proprietary to Google Cloud)	Yes (Can run in any cloud platform)	Yes (Can run in any cloud platform)
Open Source	No	No	No	Yes (Apache 2.0)

Shown in Table 4 above is a more detailed comparison of the database technologies under investigation.

4. Conclusion

When deciding on the right DBMS, there are now more options than ever and with more products are starting to adopt a multi-model approach [4]. While this paper compares the features and functionalities of these recent database technologies, it does not go in to further addressing the detail technical specifications. According to the comparative analysis Yugabyte DB shows to be feature reach with it being cloud neutral (multi-cloud) and open source as well as its fault tolerance, speed and scaling up and down without incurring down time. The traditional RDBMSs can support ACID transactions, but they cannot deliver distributed transactions or global data distribution without complicated replication architectures and Yugabyte DB incorporates these features in one [1]. But in terms of where we are in using cutting edge technology in our country Ethiopia, one of the things that also needs to be considered is not just the pure feature set of any database product or the cost of it, but also the skills on the IT team.

References

[1] Sid Choudhury, "What is Distributed SQL?", https://blog.yugabyte.com/what-is-distributed-sql/, Last accessed on November 25, 2019.

[2] Galić, Z. and Vuzem, M., "A Generic and Extensible Core and Prototype of Consistent, Distributed, and Resilient LIS". ISPRS International Journal of Geo-Information, 9(7), p.437. 2020.

[3] Wei, Jinliang. "Spanner: Google's Globally-Distributed Database." (2013)..

[4] Velmurugan, L., and S. Manoharan. "Designing Factors of Distributed Database System: A Review." Data Mining and Knowledge Engineering 12.1 (2020): 7-10.

[5] Taft, Rebecca, et al. "Cockroachdb: The resilient geo-distributed sql database." Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data. 2020.

[6] Binani, Sneha, Ajinkya Gutti, and Shivam Upadhyay. "SQL vs. NoSQL vs. NewSQL-A comparative study." database 6.1 (2016): 1-4..

[7] Orlunwo Placida, O, and Prince Oghenekaro ASAGBA. "DISTRIBUTED DATABASE MANAGEMENT SYSTEM (DBMS) ARCHITECTURES AND DISTRIBUTED DATA INDEPENDENCE." (2021).

[8] GALIĆ, ZDRAVKO. "LIS IN THE ERA OF BDMS, DISTRIBUTED AND CLOUD COMPUTING: IS IT TIME FOR A COMPLETE REDESIGN?" (2020).

[9] "What is a database?" May 2021, [Accessed 9 December 2021]. [online] Available at: https://www.oracle.com/database/what-is-database.

[10] Watt, A., Chapter 6 Classification of Database Management Systems, 2013. [Accessed 7 December 2021]. [online] Opentextbc.ca. Available at: https://opentextbc.ca/dbdesign01/chapter/chapter-6-classification-of-database-systems.

[11] Battson, D., "Top 10 considerations when choosing a Database Management system" June 2014, [Accessed 18 October 2021]. [online] Available at: https://datahq.co.uk/knowledge-hub/blog/top-10-considerations-when-choosing-a-database-management-system.

[12] Weiss, Christine. "Selecting a Database Management System (DBMS)–A Practical and Quasi-Technical Analysis of Relational Database Management Systems (RDBMS) and Object Database Management Systems (ODBMS)." University of Colorado at Colorado Springs (2012).

[13] Han, Jing, et al. "Survey on NoSQL database." 2011 6th international conference on pervasive computing and applications, pp. 363-366. IEEE, 2011.

[14] Al Hinai, A. H. "A Performance Comparison of SQL and NoSQL Databases for Large Scale Analysis of Persistent Logs.", 2016.