HiLCoE Journal of Computer Science and Technology

December 2020

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 http://www.hilcoe.net.

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

Adaptive Tool for Searching Open Educational Resources
Adem Keddi and Mesfin Kifle 1

Developing Blockchain-Powered E-Government Service to Create Transparency in Land Title Registration, the case of Ethiopia
Belay Zeleke and Mulugeta Libsie 9

Soil-Crops Suitability Analysis Modeling (SCSAM) Using Machine Learning Approach
Dan Ayana and Tibebe Beshah 17

Dictionary Based Word Sense Disambiguation for Amharic
Gashaw Sisay and Solomon Teferra 26

Security-Integrated Enterprise Architecture Framework (SIEAF)
Heven Hailu and Dereje Teferi 34

Modelling an Offline Electronic Cash Transfer Capability to the Existing Mobile Money Platform: Case of Social Cash Transfer Programs
Hussen Seid and Mesfin Kifle 43

GTCM Enabled User Involvement in eXtreme Programming Approach
Mohammed Nuru and Mesfin Kifle 51

Oilseed Leaf Disease Identification Using Image Processing
Rahel Bekele and Solomon Gizaw 59

Tigrigna Text to Ethiopian Sign Language Machine Translation
Redwan Tahir and Michael Melese 67

Near Real-time SIM Box Fraud Detection using Stream Mining
Sisay Matebe and Mesfin Kifle 73

Adaptive Tool for Searching Open Educational Resources

Adem Keddi
HiLCoE, Computer Science Programme, Ethiopia
ademkeddi@gmail.com

Mesfin Kifle
HiLCoE, Ethiopia
Department of Computer Science, Addis Ababa University, Ethiopia
kiflemestir95@gmail.com

Abstract

The advancement of technology and the popularity of the Internet created many opportunities. Educational institutions have been aware of its massive potential as a learning tool. The development of smart computing devices increased the use of online resources by the public. The Open Educational Resources (OERs) movement originated from developments in Open and Distance Learning (ODL) in the wider context of a culture of open knowledge, open source, free sharing, and peer collaboration. Nowadays, many educational and research institutions and individuals worldwide are producing and publishing OERs. There is great number of open access courseware that provide OERs freely to be used by students and anyone that is engaged in e-Learning activities.

Even though OER repositories provide open access to their databases, still there is no tool that crawls and indexes OERs that are found in different places and make the search adaptable based on the user profile making the search.

In this paper, the most appropriate approach to represent user preferences is selected. A tool that has functionalities of crawling OER databases and indexing OERs, user profiling, providing search interfaces and adapting the search results based on the user's profile is designed and developed. Finally, an experiment is conducted to test the effectiveness of the proposed tool. The tool is developed in small scale for the purpose of testing and evaluation using Open Source software. Experimentation is conducted on the developed tool to validate and verify the functional requirements and the objectives set are met.

Keywords: Personalization; Adaptive Learning; Adaptive Search Tool; Crawling; Indexing; Open Educational Resources

1. Introduction

Since the Internet was adopted and further developed by educational institutions as a means of communication in the 1970s, academics have been aware of its massive potential as a learning tool. The development of smart computing devices increased the use of online resources.

e-Learning is the use of technology to aid and enhance learning as pointed out by Carabaneanu *et al*. [1]. It can be as simple as students watching video lectures that are broadcasted from education centers or as complex as an entire university course provided online. e-Learning began decades ago with the introduction of television and overhead projectors in classrooms and has advanced to include interactive computer programs, 3D simulations, video and telephone conferencing and real-time online discussion groups comprised of students from all over the world. As technology advances, so does e-Learning making the possibilities endless.

In e-Learning, mostly Open Educational Resources (OER), are used as teaching materials. Bissell [2] described OERs as digitized materials offered freely and openly for educators, students, and self-learners to use and reuse for teaching, learning and research. OERs differ from other web resources in that they are openly licensed resources that can be used for teaching or learning. They are usually located in dedicated known repositories.

An increasing number of institutions and individuals worldwide are producing and publishing OER. Amongst these, perhaps the single most known program is MIT's Open Courseware, which provides open access to materials used to teach over 1,800 courses as pointed by Atkins *et al*. [3]. One can find

a growing list of higher education institutions participating in the creation of OER by visiting the Open Education Consortium (OEC) website [4].

OEC declared that as of July 2015 there are more than 280 member organizations worldwide. Besides MIT, University of California, University of Arizona, Cape Town University, Delft University of Technology, and Harvard University are some of the universities that made open educational resources available to the public.

At the start of the e-Learning concept, some educational institutions produced and published educational resources as additional content of the classrooms for their students [5]. After the move by MIT to put the entire course catalog online in 2001, many educational and research institutions joined the movement to make OERs open to the public.

There are too many OER repositories and courseware. These produce huge number of OERs that it is tedious to search for a particular resource in each repository separately and one by one. It is also difficult and time consuming to get a resource that matches the preferences and expertise level of the user making the search.

There are many works done on adaptable systems many of which worked on making the search in a single e-Learning system adaptable. They do not help on how to utilize resources that are found in different OER repositories and make the search adaptable to reflect the user making the query.

In this paper, we designed a tool that indexes open educational resources by crawling through OER databases to make the search in the OER repositories adaptable by making the search in the open access courseware smart enough to know the user's background and preference and hence tailoring the search results to meet the user's needs.

To achieve our objectives, we used qualitative research method to define adaptability and user preferences list in the context of web searching. Comparative study on different user profile data collection techniques is conducted to select the best method. The optimized user preferences list is developed by gathering data from different books and search engines like Google, Bing and Yahoo. This is used to create a user profile explicitly.

A tool that has functionalities of crawling and indexing open educational resources, and the selected approach to adapt the search results in open educational resources is designed and developed using the agile software development methodology. Finally, software validation and verification experiment is conducted to verify whether the tool is effective and meets its intended purpose.

2. Related Work

Many researches are done on adaptable systems and personalized learning. Greene *et al*. [6] proposed the use of anonymous profiles to develop a framework for an open source Adaptable Personal Learning Environment (APLE) by extending the already existing learning environment known as the Portland Learning Environment by adding the concept of adaptability. Their work is limited to making the existing learning environment adaptable.

Li and Chang [7] designed a personalized e-Learning system which can automatically adapt to the interests and levels of learners based on the IEEE Learning Technology Systems Architecture (IEEE LTSA), a component-based framework for general learning system. Their work is limited to adaptation of a single learning system and employs only the use of implicit information collection technique to construct user profile. This method does not guarantee the user profile created efficiently reflects the user as pointed out in [8].

After reviewing different personalization systems, Sugiyama *et al*. [9] proposed several approaches to make the web search adaptable as opposed to typical search engines which return similar results for the same query made by different individuals with different knowledge and preferences. They conducted experiments using the proposed approaches and evaluated the retrieval accuracy of the approaches to select the best one. From the experiment, the approach that utilizes user profile construction based on the modified collaborative filtering yield the best accuracy by enabling construction of more appropriate user profile and perform fine grained search that is better adapted to each user's preferences. The limitation of this work is that only implicit preferences are used which limits

the representation of the user. The use of web history for user profiling also poses a privacy concern. Another limitation is that the search will be made in the whole Internet which will return thousands of results that are not open educational resources, and filtering them will be a burden to the user.

Some researchers tried to make the search in a single e-Learning system to be adaptable. However, nowadays, there are huge numbers of OERs to be used freely for educational and other purposes. There is no previous research conducted on how to make available all of the OERs that are found in different repositories to be used by students, teachers, researchers, or anyone that embarked on e-Learning, and provide the functionality of adapting the search results. This paper used some of the findings of the above listed researches and fixed the limitations of some to propose an efficient tool that can provide adaptable search results in OER repositories.

3. The Proposed Solution

A tool that has the functionalities of indexing and crawling of OERs, constructing and managing of user profiles, and providing search results adaptation is proposed whose framework is shown in Figure 1.

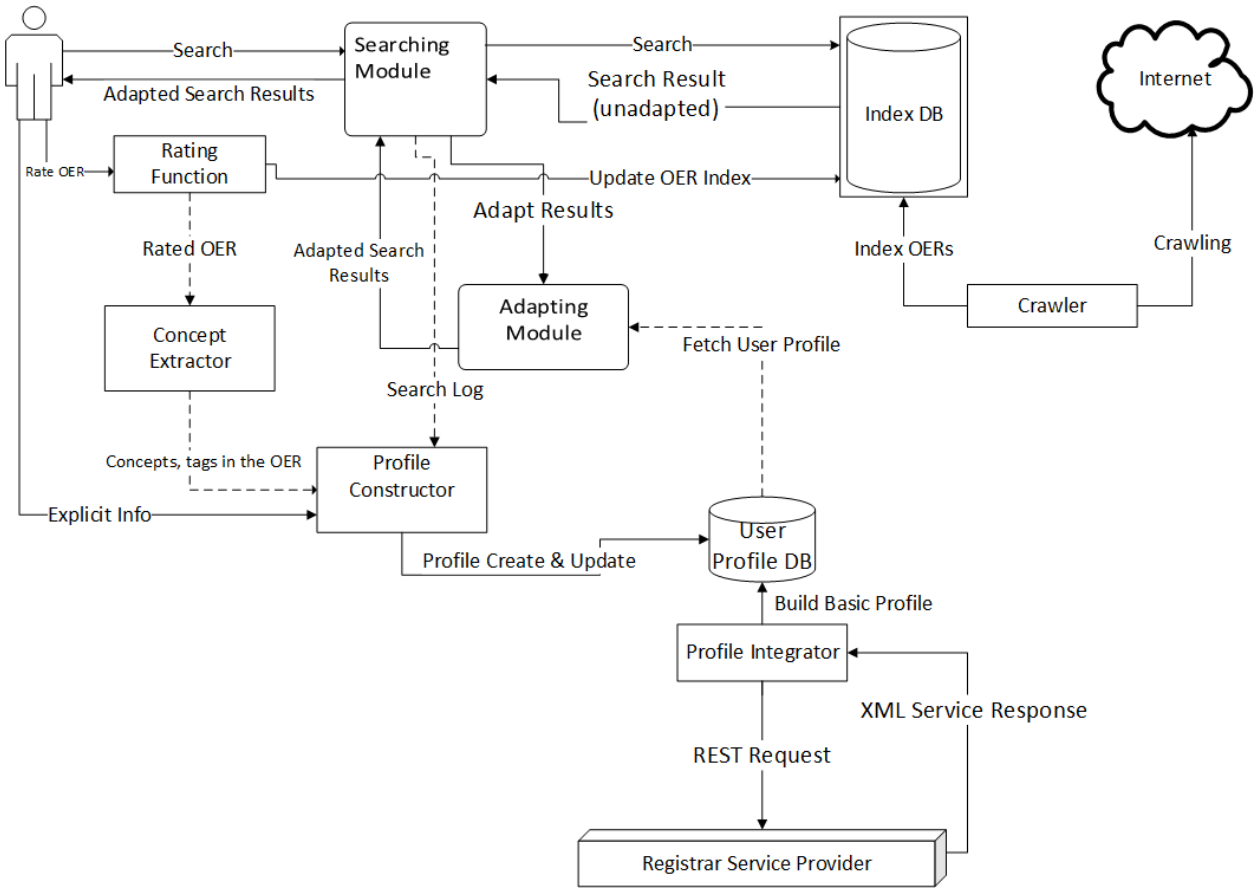


Figure 1: The Proposed System Framework with Emphasis given to User Profile Construction

As shown in Figure 1, the user searches for OERs by entering a search string in the system. The search string is passed to a controller in the search module. The controller validates the search string and performs search keyword extraction to create search keys. The search keys are then used to search in the OERs index database. Then the search module will pass the search results to the Adapting Module.

The Adapting Module retrieves the user profile from the User Profile Database and performs adaptation and returns the adapted search results to the Search Module. The search module returns the final adapted search result to the user. The user can rate OERs using the rating mechanism found in the search results page.

In the proposed tool, there are three core modules:

i. User profile construction and updating,

ii. Crawling and indexing OERs, and

iii. Searching results adaptation.

The details of each is presented below.

i. User Profile Construction and Updating

The proposed system uses user profile to adapt the search results. The following are the different clusters of data used for user model creation and management:

- *explicit preferences*: defined by the user such as sex, educational level, language skills, hobbies, topics of interest, etc.
- *implicit preferences*: that the system defines automatically by analyzing data about user behavior and activities.
- *user behavior data*: generated from the user statistics.

Among the available implicit information collection techniques, the search log is used to collect information of the user implicitly. Other techniques such as browser agents and desktop agents add some burden on clients since they are essentially client-side approaches [10]. Another advantage of using search logs is that it does not bring privacy issues to the user since it only collects information related to search.

For the user profile construction and updating, both implicit and explicit information collection methods are used to yield better result as previously conducted researches show that there is no clear answer whether implicitly created profiles are more or less accurate than explicitly created profiles [10].

ii. Crawling and Indexing OERs

The OERs databases are limited in number and the web addresses of most of them can be found from the consortium of OERs website and other related repository pages. For this application, we proposed developing a web page that contains URLs of OERs databases. If there is new digital library or OERs repository, the web address for the repository has to be added to this web page with ease. So, this web page will be used to build a URL seed list that is input to the crawler.

The high-level architecture of the crawler used for this tool shown in Figure 2 is adapted from the basic crawler cited in [11]. The basic difference is that this crawler only crawls OER repositories listed in the seed page.

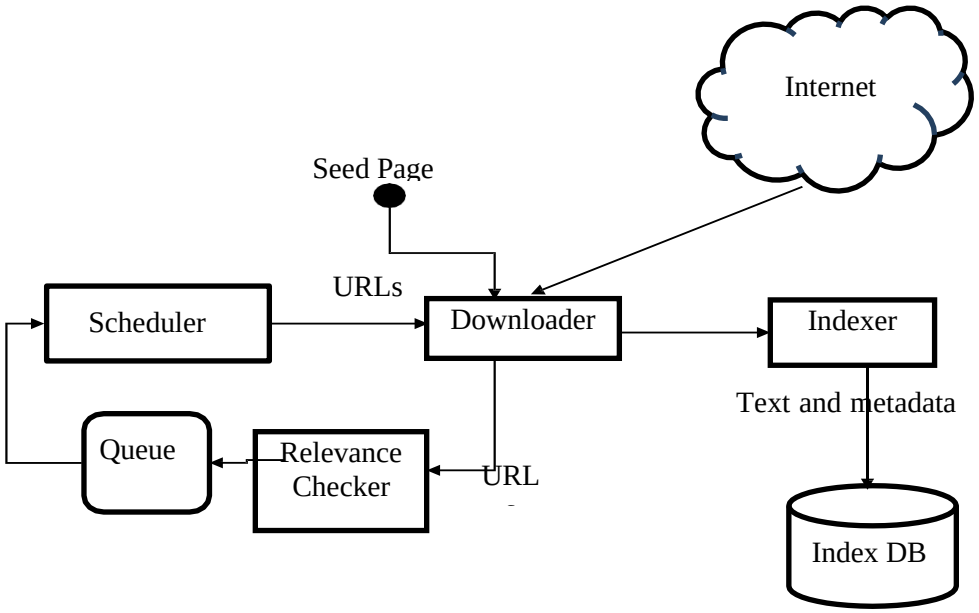


Figure 2: Adapted High Level Architecture of the Crawler

The seed page is maintained manually as there are limited number of OER repositories. The downloader gets the first URL from the seed page and downloads it. It then retrieves all the URLs in the downloaded page and passes the links to the Relevance Checker. The downloader also passes the downloaded file to the Indexer so that the file will be indexed to become suitable for searching.

The Relevance Checker checks and determines if the URL passed to it is an OER in the OER repositories. If it is determined to be an OER, the URL will be passed to the Queue and then to the

Scheduler. This process will be repeated until there are no URLs scheduled to be downloaded.

Algorithm of the Indexer

i. Get the title of the downloaded file
ii. Check if the crawled OER has keywords defined in the metadata
    a. If there are keywords in the metadata, copy them into indexed_file_keyword table
    b. Otherwise, skip
iii. Get the first page, get the first word (token), remove suffixes, capitalization, apostrophes, and put in the token list with frequency 1
iv. Get the next token, and remove prefixes, suffixes and capitalizations and compare with the list of tokens
    a. If match found, increment the frequency of the match by 1.
    b. Otherwise, add the token to the list of tokens and make its frequency
v. Go to step iv and continue until EOF is reached

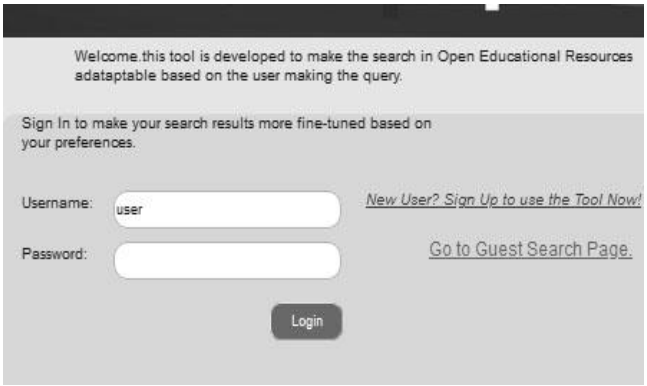


Figure 3: Login Page

4. Implementation

The system is developed in Open Source Software like MySQL Server and JavaServer Faces technology. Apache Tika library 1.2 is used for metadata extraction from the crawled resources.

The first page that is shown when the application is opened is the login page as shown in Figure 3. In this page, a user can make the search as guest user or login to the system. Logging into the system is not just for authentication purpose, but also it serves as the source of information for user profile generation in that it passes the user's ID to the system so that the necessary user profile is retrieved.

Figure 4 shows the explicit information filling or updating page.

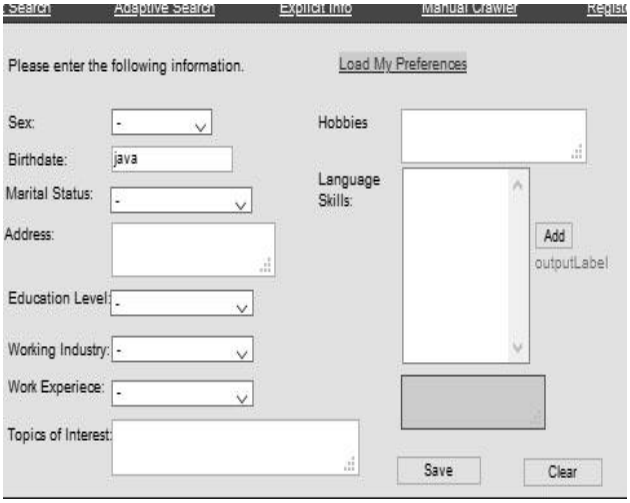


Figure 4: Explicit Information Filling Page

Figure 5 shows the search result for query string "java" made by a user with user profile of language skill of English and French and education level of graduate. So OERs that match both the language skill of the user (English and French) and education level (graduate level) are returned in the first top rows, each prioritized by their previous rates.

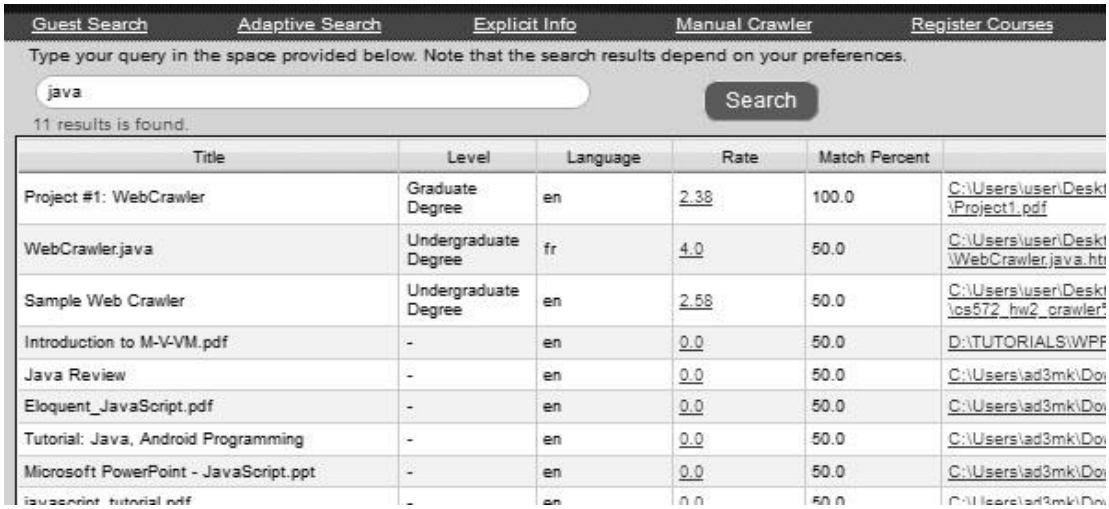


Figure 5: Search Result for Search String 'Java' for a User with Skill English and French and Educational Level of Graduate

Figure 6 shows the result of the search made by querying for "java" by a user that has only English language skill and education level of undergraduate. So OERs that are written in English and have level of undergraduate are prioritized at full matches and returned in the first rows, while OERs that match only by one of either of the two (language or education level) follow and they are prioritized by the rates given to them. Thirdly, OERs that match the search string but not the user's language and educational level will follow ordered by their rates. This ensures that all OERs that are returned by the search will be displayed to the user but adapted based on language skill, education level, and geographical location of the user making the query.

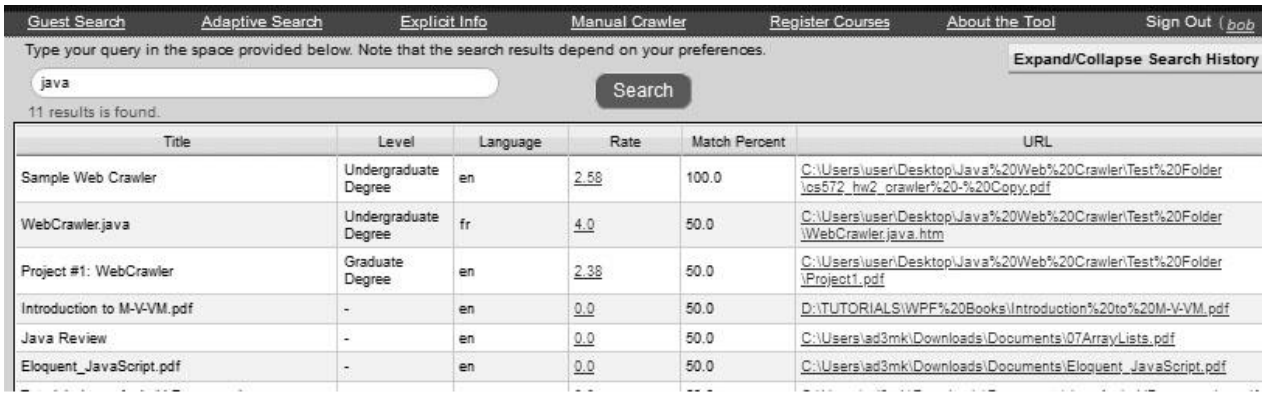


Figure 6: Search Result for Search String 'Java' for a User that Only Reads English and Educational Level of Undergraduate

For a guest user that searched for "java" without logging into the system, the search result looks as shown in Figure 7.

Note that in the search result, all are displayed without being ranked and prioritized since the system has no information about the user making the search. But here, the system can make adaptation based on the geographical location of the user which is retrieved from the IP address of the user.

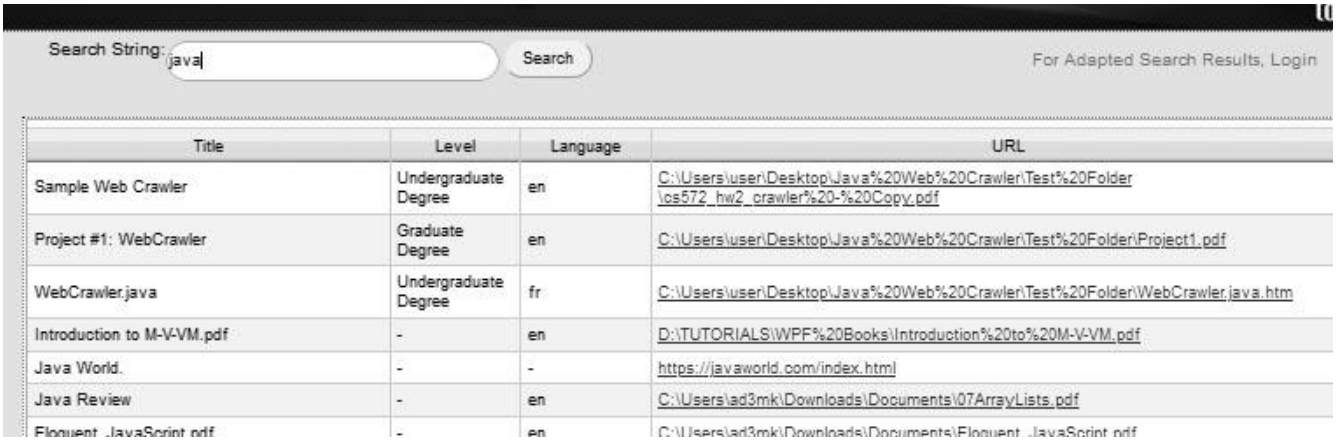


Figure 7: Search Result for Search String 'Java' for a User that is not Logged into the System

5. Software Validation and Verification

Software validation and verification is done using experimental method. For this purpose, 10 students are selected randomly with the objective to check the functional and non-functional requirements of the software and check and confirm the effectiveness of the tool in adapting search results based on the profiles of the users using the tool.

Using manual crawling functionality of the tool, few OER repositories are put into the crawler and around 300 OERs are indexed and used for the experiments.

The users have different profiles as shown below.

- Education level and field of study (Undergraduate, Graduate Computer Science, Graduate Software Engineering),
- Language they can read (e.g., English, Arabic, French, German, etc.),
- Guest users (no information about the users).

User's profile is created using the application's explicit profile collection functionality.

Out of the 10 users, 8 of them were made to perform the search in two modes: registered user with full profile and guest user. Two users used the tool as guest users only. The users are made to perform search using similar search strings. Finally, they are made to express their findings of the tool.

Those users that used the tool in the two modes found out that the results are ranked based on their language proficiency and education level filled during the explicit information collection. They also expressed the tool prioritizes the results based on the rates of the OERs. They also noted that when they use the tool in the guest mode, the tool prioritizes the search results based on the rate given to the tool.

The two users that used the tool as a guest found that the tool displays the results in unspecified order. But as the users that used the tool in both modes spotted, the search results are based on the rates given by the registered users to the OERs.

All of the users that tested the tool as registered users identified that the tool performed in conformance to their profiles and the tool's search logging functionality is promising.

6. Conclusion and Future Work

For evaluation purpose, a tool that provides all the functionalities (in small scale) is developed. Software validation and verification testing is conducted on few randomly selected users.

The main attributes used fore validation and verification are:

- the user's educational level,
- language skills,
- geographical location, etc.

From the result of the experiment, the system is verified to adapt based on the user performing the search. The evaluation of the proposed tool is done and is found to be satisfactory even though few users participated and the user profile is not mature.

The search algorithm used in this tool is basic and it should be developed more to take more user profiles in adapting and predicting the user's needs. The tool builds and updates user profiles as users

interact with the system. This is one of the outstanding features of the proposed tool.

To speak of the effectiveness of the tool in quantitative way, all of the modules will be implemented including the indexer, user profiler and user profile integrator. Evaluating the efficiency of the tool as a whole will also be done in the future.

References

[1] Luciana Carabaneanu, RomicaTrandafir, and Ion Mierlus Mazilu, “Trends in e-Learning,” in Proceedings of MMT2006, January 2006.

[2] Bissell, A. “Permission granted: Open licensing for educational resources”, The Journal of Open and Distance Learning, Vol. 24, 2009, pp. 97-106.

[3] Daniel E. Atkins, John Seely Brown, and Allen L. Hammond, “A Review of the Open Educational Resources (OER) Movement: Achievements, Challenges, and New Opportunities,” Retrieved from https://hewlett.org/wp-content/uploads/2016/08/ReviewoftheOERMovement.pdf, Last accessed on July 2015.

[4] Open Education Consortium Official website, http://www.oeconsortium.org, Last accessed on July 13, 2015.

[5] Hal Abelson, “The Creation of OpenCourseWare at MIT”, Journal of Science Education and Technology, Vol. 17, No. 2, pp. 164-174, April 2008.

[6] Green, S., Nacheva-Skopalik, L., and Pearson, E., "An Adaptable Personal Learning Environment for e-Learning and e-Assessment", in Proceedings of the 9th International Conference on Computer Systems and Technologies and Workshop for PhD Students in Computing, CompSysTech'08, 2008.

[7] Xin Li and Shi-Kuo Chang, “A Personalized E-Learning System Based on User Profile Constructed Using Information Fusion,” in Proceedings of the 11th International Conference on Distributed Multimedia Systems, DMS 2005.

[8] Liliana Ardissono, Cristina Gena, Pietro Torasso, Fabio Bellifemine, Angelo Difino, and Barbara Negro, “User Modeling and Recommendation Techniques for Personalized Electronic Program Guides”.

[9] K. Sugiyama, K. Hatano, M. Yoshikawa, and S. Uemura, “Adaptive Web Search Based on User‟s Implicit Preference”, DEWS2004, 5-B-04.

[10] Susan Gauch, Mirco Speretta, and Aravind Chandramouliand Alessandro Micarelli, “User Profiles for Personalized Information Access”, in The Adaptive Web, Methods and Strategies of Web Personalization, January 2007.

[11] Shyam Nandan Kumar, “World towards Advance Web Mining: A Review”, American Journal of Systems and Software, 2015, Vol. 3, No. 2, pp. 44-61.

Developing Blockchain-Powered E-Government Service to Create Transparency in Land Title Registration, the case of Ethiopia

Belay Zeleke
HiLCoE, Software Engineering Programme, Ethiopia
belayomo.it@gmail.com

Mulugeta Libsie
HiLCoE, Ethiopia
Department of Computer Science, Addis Ababa University, Ethiopia
mlibsie@yahoo.com

Abstract

Land is a significant resource for growth and development. The land sector is particularly susceptible to corruption and rent seeking globally. There is a high risk of corruption in Ethiopia's land administration service. Proper administration of land is critical for the future growth and development of the country. Applying information technology in the land administration service can increase the quality of service provision. A blockchain is a transparent distributed database that records details on all transactions. Once data is registered on a blockchain, it is secure and unalterable and provides a permanent record. Because of the distributed ledger nature of blockchain technology, it can increase the availability and security of a system.

This paper examines the potential impact of implementing blockchain technology on the land administration service and how it can help in creating transparency in land administration service and fighting corruption in the land sector. We implemented blockchain technology for the proposed model using hyperledger fabric and hyperledger composer fretwork. It is believed that the work contributes to an improved knowledge about blockchain technology and its potential benefits and challenges in the implementation of e-government services. We concluded from the study that blockchain powered Land Information System is the ultimate solution to create transparency, eradicate corruption, and create trust in the land administration service.

Keywords: Blockchain; E-government; Hyperledger Fabric; Hyperledger Composer; Land Information System

1. Introduction

Land is a key resource for growth and development. Access to rural land holds the promise to reduce poverty and foster sustainable development. At the same time, in the context of rapidly growing urban populations such as Ethiopia's capital, Addis Ababa, one of the fastest growing urban areas in the world, urban land and access to adequate housing becomes a highly sought after commodity. In Ethiopia, access to land is extremely important and has become a major socio-economic asset [3]. Implementing e-government in the land administration system can improve service delivery. Adopting blockchain technology for Land Information System (LIS) will increase transparency on land administration services.

The land sector is particularly susceptible to corruption and rent seeking globally. The value of land thus creates a significant opportunity for corruption on the part of those with the legal authority to assign, revoke, or restrict rights to it [1]. There is a high risk of corruption in Ethiopia's land administration. According to [2], half of all companies expect to give gifts when applying for a construction permit and paid a bribe to land administration officials. Petty corruption, tampering land owner title, and land-grabbing are common in the land administration sector. This is caused by the absence of strong institutions, transparency, clear policies, and land title registration software. This research raises the following main questions to address the problem.

1) How can blockchain technology be used to create transparency on land administration service?
2) What type of blockchain technology is appropriate for land administration service?

The general objective of this paper is to identify and develop blockchain powered e-government model for Ethiopian Land Administration service to create transparency and fight corruption.

One of the important things to fight corruption in the land administration service is transparency. Transparency refers to the availability of information to the general public and clarity about government rules, regulations, and decisions. Transparent procedures include open meetings, financial disclosure statements, freedom of information legislation, budgetary review, and audits. Transparency in government serves as a deterrent to corruption can inspire trust in the governance process. To establish trust in the government, transparent governance ensures that money is spent for the intended purpose. The government should avail the necessary information on the government's official website to create transparency [9].

Traditional database applications are based on client-server architecture. Clients are end-users of the application that request a particular service [10]. Traditional database applications have an advantage over blockchain technology in scalability, easy maintenance, and high throughput. Traditional database applications have their own limitations compared to blockchain technology. They have limitations on identifying data inconsistencies and correcting any unreliable information, robustness, security, and anybody with sufficient access to a centralized database can destroy or corrupt the data within it [10].

Blockchain is a transparent distributed database that records details of all transactions performed by the system's participant agencies. It is a decentralized public ledger that allows all participants to keep track of and verify transactions without the need for a central party [4]. Once data is registered on a blockchain, it is secure and unalterable and provides a permanent record. Blockchain provides the ultimate solution in transparency and trust, making this technology an exciting prospect for organizations that strive to conduct transparent business [13].

There are three general high-level categories for blockchain approaches: Permissionless, Private Permissioned, and Public Permissioned (Federated Blockchains or Consortium Blockchains). A consensus algorithm can be defined as the mechanism through which a blockchain network reaches an agreement. There are several consensus algorithms. The most popular are Proof-of-Work, Proof-of-Stake, Round Robin, Proof-of-Authority/ Proof-of-Identity, and Proof-of-Elapsed Time.

The blockchain platform or frameworks are software solutions that simplify the development, deployment, and support of technically complex products. Building a blockchain application from scratch can be quite complicated. However, blockchain platforms allow to develop blockchain applications with ease. Since there are a variety of blockchain platforms, we need to pick the one that meets business requirements. Here it is important to mention the different types of Blockchain platforms that can be used by enterprises. Some popular paltforms are Hyperledger Fabric, Hyperledger Sawtooth Lake, IBM Blockchain, R3 Corda, BigChainDB, and Open-chain.

Blockchain technology has a benefit in reducing transaction cost, achieving data security, creating trust in the government, creating transparency, and streamlining business processes across multiple entities (reconciliation). We believe that it is important to understand the limitations of blockchain technology and the different trade-offs that result from different choices in architecture and design. Without a clear understanding of these trade-offs, it is impossible to know where blockchain technology can be best applied, let alone whether it should be considered at all [12]. Before deciding to implement blockchain technology, one should consider the following points: Blockchain technology is not the best solution for every problem; the adoption of blockchain technology should consider infrastructure development (ICT, Electric grid, etc.); and human capacities [8]. Hence, selecting appropriate blockchain type, selecting the right consensus algorithm, and selecting the right blockchain platform are important.

2. Related Work

The work by Lazuashvili [6] is intended to evaluate Georgian land title blockchain project. The following strong factors are identified for the successful adoption of blockchain technology in the Georgian land title registration system: existing e-Government, Public-Private Partnerships, Legislative

framework, and Research and Development activities. The weakest points in the blockchain project of Georgian land title are that no specific research projects had been done before launching the pilot project in 2016 and no specific metrics were identified for the evaluation of the results of the project. The blockchain solution is not integrated into the land titling service itself, but an add-on service on the existing land title process. The issues in the current public service provision processes are identified that blockchain technology helps to reduce corruption, reducing the red tape, improving the speed of transactions, increasing data security, offering more transparent state-citizen relationships and creation of better tools of data traceability.

The purpose of the research by Corluka and Lindh [7] is to examine the implications of the implementation of blockchain technology within the real estate sector and its possible consequences. The positive impacts after implementation of blockchain are less market inefficiey, safety, more stable market, easier to trust the system, and higher transaction frequency. The negative impacts after implementation of blockchain are oligopoly, changes in work tasks, less human contact, less arbitrage opportunities, and problems with adjustments to the increased requirements. The authors draw a final conclusion that blockchain is a promising technology with great ability to fundamentally change not only the real estate market but the market overall.

The objective of the research by Oberdorf [11] is to explore the potential impact of the application of Blockchain technology in land administration and property management. The research analyzed two Ghanaian enterprises that are independently developing digital land administration platforms using Blockchain technology. The author identified the following potential impacts of blockchain platform: it has a positive impact on the perceptions of tenure security; it increases levels of transparency in land transactions; it prevents disputes; it enables access to formal credit from banks, insurances and other financial services; and it enables an active land market for buying, selling and finding investment.

One of the biggest challenges that grip various states in India is the issue of land ownership, which is partly due to the fact that there is no end-to-end effective land records management system. The issue of land ownership is so critical that almost all financial institutions heavily depend on the ownership of land-based properties for collateral security. The objective of the research by Abhishek [14] is to identify problems in the current land management system. It is also intended to evaluate the performance of blockchain-based implementation of land registration system and evaluate whether integrating blockchain technology to the current business process of land record management is a viable solution or not. Factors such as integrity of data, trust in cross-organizational work, and public verifiability are identified as reasons for the current system's problems. The author argues that to solve the current system's problems, blockchain is the best solution because using blockchain technology public verifiability can be easily provided using the cryptographic hash properties of the transactions and the hash linkage of the blocks can be used to provide the traceability of a particular record. Private and permissioned blockchains technology is a suitable cross-organizational workflow to create trust between departments. Also, the distributed architecture of the blockchains helps in preventing unauthorized modification of data via consensus mechanisms.

3. The Proposed Solution

Registering land properties and availing land information for different stakeholders have many benefits. Information technology can play a great role in the land administration process. Selecting the right technology for the land title registration system can help in creating transparency in the land administration process, track changes made on land title, and providing better service for the society. The proposed solution in this paper aims to create transparency by integrating blockchain technology in the land administration service. There are different blockchain technologies available for different industries. We choose hyperledger fabric blockchain with hyperledger composer framework for land administration service because of the following reasons [5].

- The technology has the capability to track the origin of every transaction inside the blockchain ledger,
- Security,
- Support for the participation of multiple organizations,
- The simplicity of the technology,
- API support,
- Low implementation and running cost,
- Participants identification,
- High transaction throughput performance, and
- Low latency of transaction confirmation.

Hyperledger composer is a framework to accelerate the development of applications built on top of Blockchain platforms such as hyperledger fabric. Hyperledger Composer is used to model business network quickly, containing existing assets and the transactions related to them. Assets are tangible or intangible goods, services, or property. As part of our business network model, we define the transactions which can interact with assets. Business networks also include the participants who interact with them, each of which can be associated with a unique identity, and across multiple business networks. Hyperledger composer uses model file, scrip file, Access Control file and Query file to build blockchain application. The model file is used to model asset, participant, transaction, and event. The script file is used to implement transactions, in our case it is used to transfer landowner title. The access control file is used to define access control rules. Access control rules allow fine-grained control over what participants have access to what assets in the business network and under what conditions [5].

The proposed architecture has three layers: client layer, middle (hyperledger composer) layer, and Blockchain Layer (Hyperledger fabric).

The *client layer* requests service from blockchain network through the middle layer (hyperledger composer). The client layer can be an existing legacy system or external third party application or playground.

The *middle layer* (Hyperledger Composer) contains API, Hyperledger composer REST server, and SDK. API is used to integrate with existing land information systems and requests service from the blockchain network. Integrating existing land information systems allows to pull data from land information systems and convert them to assets or participants in a Composer business network. The REST Server will generate a set of REST APIs from a LIS Model. REST API can then be called from a LIS to interact with a Blockchain. When the LIS calls one of the REST APIs, the REST server will then submit a transaction to the hyperledger fabric blockchain. The SDK provides APIs for transaction processing, event handling, node traversal, and membership services. Land information system can communicate with hyperledger fabric network using REST API. REST API communicates with hyperledger composer REST server and REST server communicates with hyperledger fabric network. Using Playground locally, we can connect to a Web Browser which works in the browser's local storage, or we can use Connections to a real fabric usually in a group.

The *third layer* is hyperledger fabric. It contains CA (Certificate Authority), channels, peers, and chaincode. Each organization has a CA to authenticate blockchain network users. The channels are used to ensure private and confidential communication between privileged organizations' nodes. Channels help to partition the fabric exchange confidential private transactions with the isolation of traffic between municipalities (members) and establish data privacy by separating ledgers per channel. We can also create transparency on the land administration service by creating channels and expose land information to the public or stakeholders or the society for auditing the land information that can help to create transparency and eradicate corruption on land administration. We can anchor peer nodes on a channel that all other peers can discover and communicate with. Each member (organization) on a channel has at least one anchor peer (on the proposed architecture we recommend at least two anchor peers to prevent a single point of failure), allowing peers belonging to different organizations to discover all existing peers on a channel. A peer receives ordered state updates in the form of blocks from the ordering service and

maintains the state and the ledger. Peers can additionally take up a special role as an endorsing peer, or an endorser. The orderer node is responsible for packaging transactions into blocks and distributing them to anchor peers across the network. Multiple organizations can join the hyperledger fabric network but for the purpose of demonstration, we are showing only two organizations in the architecture shown in Figure 1.

In the proposed system, persistent data is stored in non-volatile storage such as hard disk. LIS uses a distributed blockchain ledger to store persistent data. The ledger component at each peer maintains the ledger and the state on persistent storage and enables simulation, validation, and ledger-update phases. Broadly, it consists of a block store and a Peer Transaction Manager (PTM). The ledger block store persists transaction blocks and is implemented as a set of append-only files. Since the blocks are immutable and arrive in a definite order, an append-only structure gives maximum performance. In addition, the block store maintains a few indexes for random access to a block or to a transaction in a block. The PTM maintains the latest state in a versioned key-value store. It stores one tuple of the form (key, val, ver) for each unique entry *key* stored by any chaincode, containing its most recently stored value *val* and its latest version *ver*. The version consists of the block sequence number and the sequence number of the transaction (that stores the entry) within the block. This makes the version unique and monotonically increasing.

The PTM uses a local key-value store to realize its versioned variant, with implementations using LevelDB and Apache CouchDB. PTM provides a stable snapshot of the latest state to the transaction. The ledger component tolerates a crash of the peer during the ledger update as follows. After receiving a new block, the PTM has already performed validation and marked transactions as valid or invalid within the block. The ledger now writes the block to the ledger block store, flushes it to disk, and subsequently updates the block store index. Then the PTM applies the state changes from write all valid transactions to the local versioned store. Finally, it computes and persists a value savepoint, which denotes the largest successfully applied block number. The value savepoint is used to recover the indexes and the latest state from the persisted blocks when recovering from a crash.

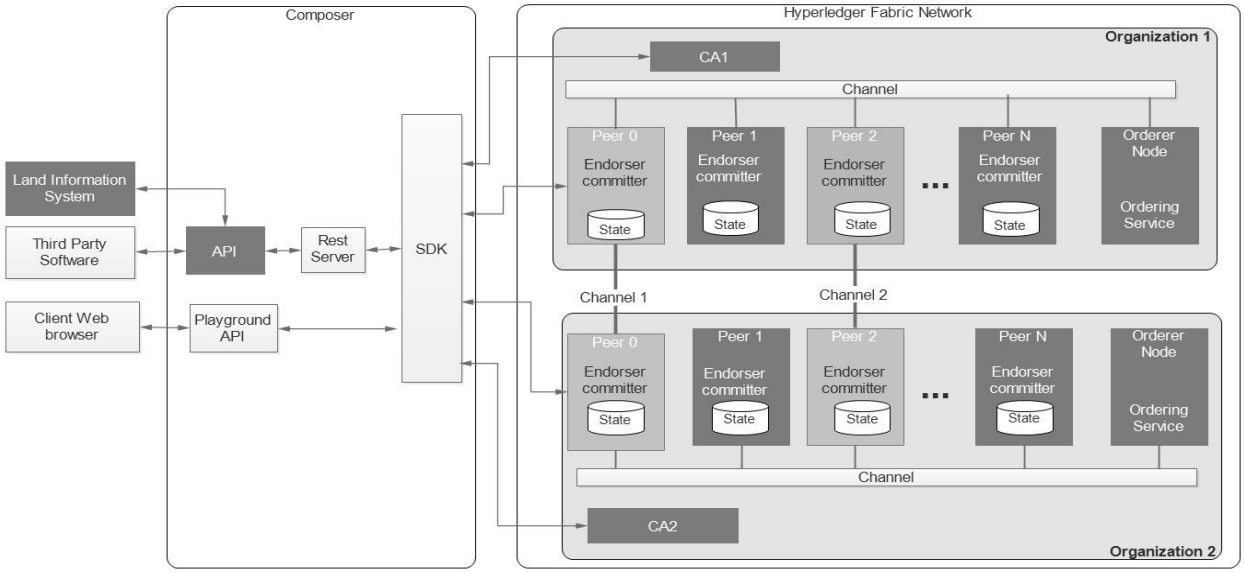


Figure 1: Hyperledger Composer Base System Deployment Architecture

For the production environment disaster recovery is recommended, persistence is part of a high availability suite which provides replication and instant failover. In all cases, to protect against data corruption, it is recommended to take regular backup of the storage associated with every deployed component. Because the ledger is shared across all the peers and ordering nodes, taking regular backups is critical. For example, if incorrect data is accidentally propagated to a peer's ledger or if data is

mistakenly deleted, the incorrect data might spread to the ledgers of some other peers. This would require the restoration of the ledgers of all the peers from an established backup point to ensure synchronicity. We can decide how often to perform the backups based on our recovery needs, but a general guideline would be to take daily backups.

4. Implementation

Implementation of the land registry system did not include all functionalities identified. For the purpose of demonstration, we considered only the basic functionalities. For the development of blockchain-based LIS, we used hyperledger fabric, hyperledger composer, and hyperledger composer playground as development tools. To run Hyperledger Composer and Hyperledger Fabric on a virtual box machine, it is recommended to have at least 20 GB of hard disk space and 2 GB RAM on the host machine. The following should be installed: Docker Engine, Docker-Compose, Node, npm, git, and Python. The following tools should be installed for development, testing, and run the application: Command Line Interface, Generator Hyperledger composer, Yeoman code generator, Playground Web User Interface, and Hyperledger Fabric. We can control our runtime using a set of scripts which can be found in the development folder *~/fabric-dev-servers*. The first time we start a new runtime, we need to run the start script, then generate a PeerAdmin card using the *./createPeerAdminCard.sh* script. We can start and stop our network runtime using *~/fabric-dev-servers/stopFabric.sh*, and start it again with *~/fabric-dev-servers/startFabric.sh*.

a. Creating Business Network, Defining Asset, Participant, Transaction and Access Control

We will walk through setting up a LIS blockchain network, defining our assets, participants, and transactions, and testing our network by creating some participants and an asset, and submitting transactions to change the ownership of the land title from one landowner to the other.

Creating a New Business Network (LIS Blockchain Network): To create a new business network from scratch we set basic information of LIS blockchain such as network name, description, and network Admin card. After providing all requirements, we can deploy the LIS blockchain network. Now that we have created and deployed the LIS blockchain network, we can see a new LIS blockchain network card called *admin* for our LIS blockchain network *blockchain-land-info* in the wallet. LIS blockchain network cards (wallets) combine a connection profile, identity, and certificates to allow a connection to a LIS blockchain network in Hyperledger Composer Playground. The wallet can contain LIS blockchain network cards to connect to multiple deployed LIS blockchain networks. Playground allows us to create and edit model, script, access, and query files that are used to define LIS blockchain networks. We have two model files. The first one is base model file, and the second one is land model file. Base model file is used to model concept, enumerated types (enum is predefined data type) and abstract participant. Land model file is used to define the assets, participants, transactions, and events in our LIS blockchain network.

Define Participant: Participants are members of the LIS blockchain network. To capture participant information on the LIS blockchain network, we should define first on the model file. A model file like a database in traditional applications is used to define data filed on the blockchain network. In the LIS blockchain network, there are six participants, two assets and one transaction. The common data fields are defined in the concept, abstract participant and enumerator in the model file. Assets are the resources which our business model has. Land and certificates are assets in our model. Creation, deletion, and update assets are supported by default by Composer. The way that the participants interact with the network and with the assets is through transactions which are the chaincode functions with certain parameters invoked by a user of the network. Any transaction invocation is always recorded in an immutable historical record. All the creation, deletion and update transactions are already supported by default by Composer, but all other behaviors need to be modeled, and sometimes, even these default transactions should be made restricted and replaced by custom made ones, to ensure system integrity.

Transaction: Now that the domain model has been defined, we can define the transaction logic for the LIS blockchain network. Composer expresses the logic for a LIS blockchain network using JavaScript functions. These functions are automatically executed when a transaction is submitted for processing. We created a script file and write chainofcode (smart contract) to perform a transaction (transferring land title from one land owner to a third person).

Access Control: Access control file is used to define access control rules for the LIS blockchain networks. Our network is simple, so the default access control file does need editing to add a participant's role. The basic file gives the current participant admin (NetworkAdmin) full access to the LIS blockchain network and system-level operations. Based on LIS access control rule we can modify access control file to restrict the role of the participant in the LIS blockchain network.

Now that we have models, script, and access control files, we need to deploy and test our LIS blockchain network. Hyperledger composer playground has a feature to deploy business network.

b. Testing LIS Blockchain Network

We need to test our LIS network by creating some participants (in this case *land Owner*, *Land Owner Family*, and *Land Administrator*), creating an asset (Land), and then using Land title transfer transaction to change the ownership of the land

Create Participant (Landowner): We open Landowner from Participant Menu from the composer playground, check the optional checkbox to insert optional properties of a landowner, and insert land owner detail information. Then the system automatically evaluates and adds landowner information to the LIS blockchain network. After creating landowner on the network we can update, delete and view landowner information in the same way we can create, update and delete Land Administrator, Land Registrar, System User, and landowner family.

Creating Asset (Land title): Another integral component of the system is the assets for the asset management aspect of land registration. If traceability of the land is to be achieved, these products need to be modeled as assets, so that the network can register which ones exist, their status, and any changes that happen. We can create new land information on the LIS blockchain network by opening Land registration page from Asset Menu from the home page of playground. To insert optional properties of Land we check on the optional properties checkbox and insert land detail information as shown in Figure 2. The peer node evaluates and adds to the LIS blockchain network. We can also update and delete land information after land information is created on the LIS blockchain network.

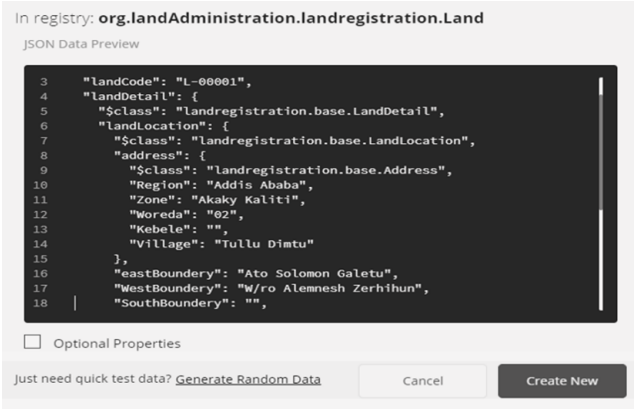


Figure 2: Creating New Asset

Transaction (Transfer Land Title): The historian is a specialized registry which records successful transactions, including the participants and identities that submitted them. The historian stores transactions as Historian Record assets, which are defined in the Hyperledger Composer system namespace. To transfer Land title, we open a transaction page and select transaction type "LandTransfer". Insert Landcode and new owners ID and submit the transaction, then the system automatically evaluates and changes the owner of land title. Once a transaction is created on the LIS blockchain network, the transaction history will not be removed from the ledger. This feature helps in the creation of transparency and helps to track transaction history. We can view transaction history by clicking on the *All Transaction* button under the transaction submenu and the transaction list will be displayed on the page. It contains Data Time, Entry Type, and Participant columns. We can also view transaction detail of each record.

5. Conclusion and Future Work

The land sector is particularly susceptible to corruption and rent seeking globally. The value of land thus creates a significant opportunity for corruption on the part of those with the legal authority to assign, revoke, or restrict rights to it [1]. There is a high risk of corruption in Ethiopia's land administration. One of the important things to fight corruption in the land administration service is transparency. Transparency in government serves as a deterrent to corruption and can inspire trust in the governance process. Blockchain is a transparent distributed database that records details of all transactions performed by the system's participant agencies. It is a decentralized public ledger that allows all participants to keep track of and verify transactions without the need for a central party. Blockchain technology has a benefit in reducing transaction cost, achieving data security, creating trust in the government, creating transparency, and streamlining business processes across multiple entities (reconciliation).

This paper examined the potential impact of implementing blockchain technology on land administration service and how it can help in creating transparency and fight corruption in the sector. We developed blockchain-powered Land Information System using hyperledger composer playground to simulate blockchain technology.

Future work will be further study on blockchain technology, develop mobile and web applications and integrate them with the blockchain application.

References

[1] J. Plummer, "Diagnosing Corruption in Ethiopia", World Bank, 2012.

[2] "Ethiopia corruption report", August 2017, https://www.business-anti-corruption.com/ country-profiles/ethiopia/, Last accessed on Feb 6, 2019.

[3] S. Lindner, Transparency International, "Overview of corruption in land administration", U4 Expert Answer, 2014, pp. 1-9.

[4] Jefrey Dinham, "A practical guide to how blockchain could end corruption in South Africa", Feb 16, 2019, https://qz.com/africa/1238247/corruption-in-south-africa-blockchains-open-ledger-could-bring-transparency

[5] "Hyperledger composer", retrieved from https://hyperledger.github.io/composer/latest/, Last accessed on November 12, 2019.

[6] N. Lazuashvili, "Integration of the Blockchain technology into the land registration system.A case study of Georgia", Master's thesis, Tallinn University of Technology, 2019.

[7] D. Corluka and U. Lindh, "Blockchain A new technology that will transform the real estate market", Master's thesis, Royal Institute of Technology, Stockholm, Sweden, 2017.

[8] R. Zambrano "Unpacking the disruptive potential of blockchain technology for human development", International Development Research Centre, Ottawa, Canada, August 2017, https://idl-bnc-idrc.dspacedirect.org/ bitstream/handle/10625/56662/IDL-56662.pdf

[9] "What is Transparent Governance?", https://market-research-tools.knoji.com/what-is-transparent-governance/, Last accessed on June 21, 2019.

[10] Shaan Ray, "Blockchains versus Traditional Databases", https://towardsdatascience.com/ blockchains-versus-traditional-databases-e496d8584dc, Last accessed on Nov. 23, 2020.

[11] V. Oberdorf, "Building Blocks for land administration, The potential impact of Blockchain-based land administration platforms in Ghana", July 2017.

[12] G. Hileman and M. Rauchs, "Global blockchain benchmarking study", Cambridge Center for Alternative Finance, University of Cambridge, 2017.

[13] CMS Media, "Canada using blockchain for transparent administration of Government contracts", https://www.cms-connected.com/ News-Archive/August-2018/Canada-Using-Blockchain-for-Transparency-of-Government-Contracts, Last accessed on Jan 29, 2019.

[14] G. Abhishek, "Property Registration and Land Record Management via Blockchains", Master's Thesis, Indian Institute of Technology, Kanpur, June 2019.

Soil-Crops Suitability Analysis Modeling (SCSAM) Using Machine Learning Approach

Dan Ayana
HiLCoE, Computer Science Programme, Ethiopia
Engineering Corporation of Oromia
danpro.it@gmail.com

Tibebe Beshah
HiLCoE, Ethiopia
School of Information Science, Addis Ababa University, Ethiopia
tibebe.beshah@gmail.com

Abstract

For a country whose economy is strongly dependent on agriculture, conducting land and environmental study and assessing the land's physical nature, chemical characteristics, and topographic formation are critical for improvement. While we are in the era of AI (Artificial Intelligence) where most of the industries are actively using ML (Machine Learning)/Data Mining (DM), Soil Suitability Classification (SSC) remains one of the tedious manually crafted and resource-intensive activities in the agricultural sector.

Thus, to enhance these processes, following Design Science Research Methodology and Cross-Industry Standard Process for Data Mining (CRISP-DM) design and development frameworks, a study was made experimenting with classification algorithms of Rapidminer, WEKA, and Python DM/ML libraries.

In the Rapidminer tool, experiments were done on various ML algorithms using overall crops, Six Crops, and Crop-Based datasets. Prediction accuracies of 94.7%, 96%, and 99% were achieved using Naïve Bayes, Generalized Linear Model, and Gradient Boosted Tree algorithms, respectively. Similarly, in the WEKA tool, prediction accuracy of 96% was found on the J48 tree algorithm. In Python, using a crop-based dataset, different experiments were done with the default and applying the Oversampling technique to minority classes. The algorithms' generalization capability was also evaluated with 10-fold cross-validation and other evaluation matrices.

From the experiments, selecting the top-scored algorithms (decision tree, random forest, extra tree, Multilayer Perception (MLP), and K-Nearest Neighbor (KNN)) and settings, the Soil-Crop Suitability Prediction Prototype was created for six crops and evaluated with user data. In the evaluation, professionals confirmed that the predicted result by the prototype was 88%-97% matching with manually classified crop suitability data. Likewise, in their feedbacks, professionals indicated that the SCSAM (Soil-Crops Suitability Analysis Modeling) artifact is an opportunity for soil suitability analysis to minimize experts' efforts and resources.

Keywords: Design Science Research; CRISP/DM; Soil Suitability; Classification

1. Introduction

For a country whose economy strongly depends on agriculture, land and water resource utilization is very crucial for the benefits of its citizens [1].

Over the past two to three decades, the rapid population growth, frequent technological advances, industrial development, and agricultural expansion have led to inequality and imbalances in the supply of adequate water, energy, and other basic resources to inhabitants.

From environmental resources, land is used as the basis for agriculture, industry, settlement, and other socio-economic development activities.

Knowing the soil's physical nature, chemical characteristics, and topographic formation helps to understand its potential and improve productivity. Land and environmental study helps to ensure long-term productivity and sustainable development through the appropriate use of land resources [2]. Thus, in this study, Land Suitability Assessment (LSA) is used to find where suitable land is and to identify its potential as to what kind of uses it is best suited [3].

Soil suitability analysis and assessment are usually done through the determination of the most important factors such as availability of nutrient elements, pH, cation exchange capacity, oxidation-

reduction conditions, water holding capacity, and the like [4] that have a direct impact on soil properties. In soil suitability assessment, professionals are expected to apply Soil Science knowledge in the collection, mapping, analysis, testing, and interpretation process.

Although results often depend on several factors, the evaluation and interpretation processes are not an effortless task [5]. It requires multidisciplinary and cross-sectoral approach to contemplate the biophysical, socio-economic, and environmental characteristics. As a result, it consumes too much time and resources and it is highly vulnerable to personal opinions.

Yet, there are Artificial intelligence (AI) and Machine Learning (ML) technology-based applications on the market to automatically learn from data and improve from experience without being explicitly coded [6]. Banks, government offices, service-giving sectors, insurances, etc. apply those technologies to their problems and have begun to reap the benefits. However, land-use study and soil analysis continued as the most time taking and expensive activities in the agricultural sector.

In Ethiopia, especially in Oromia and surrounding regions, the Regional State Land Administration Bureaus collaborating with different NGOs funded many Integrated Land-Use Plan (ILUP) studies for major rivers and sub-basin lands [7]. However, project owners and professionals have not still started recognizing ML to estimate and forecast similar ILUP projects. Instead, experts have followed common steps and procedures for similar land-use study projects that often lead to delays and extra budgets.

Thus, the objective of this paper is to investigate soil suitability analysis and build an automated knowledge-based Soil-Crop Suitability Classification Prototype that can enhance suitability analysis and the interpretation processes of land-use study.

2. Related Work

Land evaluation for crop production has provided useful information to stakeholders to improve crop yield and soil management.

The work in [9] used a Parallel Random Forest (PRF) classifier for an ML prediction model for the Sorghum crop. The authors set-up experiments using PRF, Linear Regression (LR), Linear Discriminate Analysis (LDA), KNN, Gaussian Naïve Bayesian (GNB), and Support Vector Machine (SVM). 10-fold cross-validation was used to evaluate performance. According to their findings, PRF has a better accuracy result of 96%.

The work in [10] used ML techniques to predict yield by analyzing soil datasets. The crop yield prediction problem was formalized as a classification rule where Naïve Bayes and KNN methods were used.

Similar research was conducted on crop suitability classification to improve the LIMEX system [11]. The study used a hybrid approach (c4.5 tree) considering 11 attributes.

A newly proposed ensemble classifier generation technique called RotBoost, which is a combination of Rotation Forest and AdaBoost, is used for land suitability classification in [8]. The researchers found that RotBoost generates ensemble classifiers with significantly higher prediction accuracy.

So far, this research was focused on the prediction of the suitability of the soil to the required crops that are grown in most areas of the country and has complete features. The main difference or feature of this study is that, unlike [9] and [10], we have included most of the determinant chemical features and used more than 22,000 soil sample data collected from the different agro-ecological zones of Ethiopia.

3. The Proposed Solution

In order to automate and enhance the soil suitability analysis work for medium to large scale land use planning, we have made a systematic investigation using the design science research methodology [12] and Cross-industry standard process for data mining (CRISP-DM) framework. Afterward, an attempt was made to develop ML-based soil suitability prediction model.

Our proposed method to predict Soil-Crop suitability is through the implementation of supervised classification machine learning algorithms using Decision tree, Random forest, Extra tree, SVM, and KNN classifiers.

After understanding the business process, essentials, and challenges, important data were collected from Oromia Water Works Design &

Supervision Enterprise (OWWDSE), Land-Use Planning Study and Environmental Protection Subprocess Division. The collected raw data are the soil survey data based on a 1:50,000 mapping scale for over 41 million hectares of land representing more than 50,000 soil mapping units from Oromia, Somali, Gambela, and Afar regions.

For a machine to learn the problem and make predictions effectively, data preprocessing and feature engineering are the fundamental processes. Therefore, to transform raw data into important features, a hybrid feature selection technique (Knowledge Expert Feature Selection and Traditional Attribute Selection Method) was implemented. First, to reduce and keep only very determinant attributes, Knowledge Expert Feature Selection method [13] based on expert consultation was used. Then those identified features are further preprocessed using traditional attribute selection methods found in each ML tool.

Discussing with experts, the possible soil-crop suitability attributes from physical, environmental, and chemical features are identified. Table 1 shows suitability determinant chemical features that are considered in the soil suitability analysis for the required crop.

Table 1: Soil Suitability Determinant Chemical Features

No.	*Chemical properties*	*Code*
1.	Cation Exchange Capacity	CEC
2.	Base saturation percentage	BS
3.	Electrical conductivity	EC
4.	Exchangeable Calcium	Ca
5.	Exchangeable Magnesium	Mg
6.	Exchangeable Potassium	K
7.	Exchangeable Sodium	Na
8.	Exchangeable sodium percentage	ESP
9.	Total Nitrogen	TN
10.	Organic Matter	OM
11.	Organic Carbon	OC
12.	Soil Reaction	PH
13.	Available Phosphorus	AvP
14.	Carbon to Nitrogen Ratio	C: N
15.	Exchangeable Calcium to Exchangeable magnesium ratio	Ca: Mg
16.	Calcium carbonate	Caco3
17.	Potassium Chloride	KCL

Along with that, non-relevant attributes and a large number of missing values were also discarded from the data through the Listwise or case deletion method [14]. In the meantime, the attribute values measured in different parameters were uniformly assigned with similar values. Inconsistently labeled attribute values were replaced with consistent labeling. Besides, the data details that differ significantly from other observations are handled in the standard ways [15].

For a machine-learning algorithm to improve its prediction performance and to reduce overfitting, selecting the best features among all the data is very essential [16]. Therefore, from available attribute selection methods, in the RapidMiner, the Remove Correlated method was used to clean above 0.9 correlated attributes and found a total of 9 nominal and 19 numeric attributes. Similarly, in WEKA, "Info Gain attribute Eval" filter method was used to rank the attributes and to remove those with score below the threshold. In the process, 9 nominal and 20 numeric attributes were found. In Python using Sklearn.feature_selection, the attribute Variance Thresholds are evaluated. The possible determinant number of attributes "K" was identified and using the SelectKBest method, only very important features (K-number of features) were selected.

After applying data preprocessing and preparation techniques for the collected ILUP data, we got three dataset options with 22,532 rows of soil-related processed datasets. Then for further processing and experiments, the generated datasets were transformed into the model experimentation phase to train and test the ML algorithms.

4. Experiments and Results

Since the right model can only be identified through experimentation, different classification algorithms and settings in the RapidMiner, WEKA, and Python were tested.

4.1 Experiment using RapidMiner DM/ML

Cleaning and preprocessing the raw data using data cleansing feature and removing 0.9 correlated attributes, experiments were made using preprocessed dataset. In the experiment, different classification algorithms, prediction accuracy, and the performance of a learning model were tested

using ILUP soil datasets. Among the experimented five algorithms using OverAllCrops, the best accuracy score of 94.7% was obtained in the Deep learning model. In the second dataset option using the Six Crops dataset, the best score was found on Generalized Linear Model and Deep Learning algorithms.

For crop-based experiments, applying auto and dummy encoding to the categorical attributes, two types of experiments were made for Maize and Onion crops. From the tested algorithm and settings, Naïve Bayes, Generalized Linear, and Gradient Boosted Trees algorithms show a better score in the dummy encoded method. Table 2 shows the experimented algorithms result in terms of accuracy.

Table 2: RapidMiner Experiment Summary

Algorithm	*OverAll Crop*	*SixCrop*	*Maize Auto*	*Onion Auto*	*Maize Dummy*	*Onion Dummy*
Naive Bayes	93.9%	92%	84%	85.48 %	90.2%	90.2%
Generalized Linear Model	-	96%	97%	97.2%	98.2%	99.02%
Logistic Regression	-	94%	98%	98.90 %	97.80%	98.90%
Fast Large Margin	-	Error	72%	85.1%	95%	95.59%
Deep Learning	94.7%	95%	98%	99.05 %	98.2%	98.99%
Decision Tree	11.4%	40%	87.00 %	93.72 %	79.3%	68.15%
Random Forest	18.5%	36.00%	72.00 %	85.40 %	72%	65.56%
Gradient Boosted Trees	-	Error	99%	99.1%	97%	99.18%

4.2 Experiment using WEKA

The raw data were preprocessed using the "Info Gain attribute Eval" filtering method and 9 nominal and 20 numeric attributes were prepared. Then, using available classification algorithms or learning schemes, different soil suitability experiments were made with the preprocessed datasets. To find out the most suitable algorithm, multiclass supporting classification algorithms were tested. Meanwhile, to rank their performance, Cross-validation, K-fold, and 70:30 training test splitting techniques were used.

Table 3: Weka based Classification Accuracy Result

Algorithm	*Validation*	*Overall Crop*	*Six-crops*	*Crop based*					
				Maize	*Onion*	*Rice*	*Sesame*	*Sugarcane*	*Tomato*
Naive Bayes Updatable Auto encoding	10-fold	73.89%		71.44%		-	-	-	-
	70:30:00	73.22%	-	71.88%	-	-	-	-	-
Naive Bayes Updatable Nominal to Binary encoding	10-fold	-	-	-	-	-	-	-	-
	70:30:00	-	-	-	-	-	-	-	-
Decision Tree	10-fold	96.27%	97.00%	98.14%	98.72%	95.41%	97.31%	99.81%	99.65%
	70:30:00	95.86%	96.64%	97.97%	98.73%	98.73%	97.09%	99.69%	99.69%
J48	10-fold	97.02%	97.46%	98.69%	98.86%	98.86%	98.21%	99.47%	99.49%
	70:30:00	96.60%	97.10%	98.76%	98.77%	98.77%	98.00%	99.33%	99.36%
Random Forest Classification	10-fold	96.36%	97.04%	98.44%	98.75%	98.75%	97.29%	99.87%	99.69%
	70:30:00	96.01%	96.49%	98.17%	98.55%	98.55%	96.98%	99.81%	99.67%
GaussianNB	10-fold	90.28%	-	-	-	-	-	-	-
	70:30:00	86.57%	-	-	-	-	-	-	-
K- Nearest Neighbors Algorithm	10-fold	69.45%	-	84.06%	-	-	-	-	-
	70:30:00	-	-	-	-	-	-	-	-

Using Auto and nominal to binary encoding methods, two types of experiments were made with the OverAllCrop dataset. The experiment result indicated that auto encoder has a better score than Nominal to binary encoding. From the tested algorithms, an accuracy score of >95% was found on Trees.J48, Decision Tree Classifier, Naïve Bayes, and Random Forest classifier in both 10-fold cross-validation and 70:30 splitting methods. Similarly, an experiment was made with SixCrop and the crop-based dataset and a better score was found with Trees.J48, Decision Tree Classifier, and Random Forest algorithms in the K-folding cross-validation and training test split techniques. The summary results are listed in Table 3.

4.3 Experiment using Python with its ML/DM Library

In contrast to the two data mining tools, to preprocess raw data, as a programming language, Python's ML libraries need to be coded. Therefore, further data processing and suitability experiment were made using a crop-based dataset. In the preprocessing, outlier values were handled using interquartile and replace methods. Using the round() function, numerical values were uniformly set to have only 3 digits. Empty and inconsistent value fixing code was written.

Since all ML models are mathematical models, the categorical values in the dataset need to be converted into numeric values. Alhough Python provides categorical to numeric encoding techniques [17], OneHotEncoding, oneLabelEncoder and Binary Encoding were not able to assign similar values in training test and user data provision cases. Therefore, to handle this issue, we wrote a categorical data handling code and were able to replace an identical numeric value with similar objects in all cases.

To select the best determinant features from the dataset, using the Sklearn Variance Threshold class parameter (Threshold = .51), the maximum number of determinant attributes K = 27 were identified. To identify the K number of attributes in each crop, SelectBest 'K' was implemented [18]. In each crop, the number of best determinant features were identified as follows:

- Maize [SMU, Major_Soil, Texture, Drainage, Depth, Slope, Elevation, PH_Water, PH_Kcl, EC, Na, K, Ca, Mg, CEC, BS, TN, OC, OM, C_N, P2O5, CaCO3, ESP, Ca_Mg, K_Mg, Ca_Mg_k, SUM]
- Onion [SMU, Major_Soil, Drainage, Depth, Slope, Temperature, Elevation, PH_Water, PH_Kcl, EC, Na, K, Ca, Mg, CEC, BS, TN, OC, OM, C_N, P2O5, CaCO3, ESP, Ca_Mg, K_Mg, Ca_Mg_k, SUM]
- Rice [SMU, Major_Soil, Texture, Drainage, Depth, Slope, Temperature, Elevation, PH_Water, PH_Kcl, EC, K, Ca, Mg, CEC, BS, TN, OC, OM, C_N, P2O5, CaCO3, ESP, Ca_Mg, K_Mg, Ca_Mg_k, SUM]
- Sesame [SMU, Major_Soil, Drainage, Depth, Slope, Temperature, Elevation, PH_Water, PH_Kcl, EC, Na, K, Ca, Mg, CEC, BS, TN, OC, OM.3, C_N, P2O5, CaCO3, ESP, Ca_Mg, K_Mg, Ca_Mg_k, SUM]
- Sugarcane [SMU, Major_Soil, Drainage, Depth, Slope, Temperature, Elevation, PH_Water, PH_Kcl, EC, Na, K, Ca, Mg, CEC, BS, TN, OC, OM, C_N, P2O5, CaCO3, ESP, Ca_Mg, K_Mg, Ca_Mg_k, SUM]
- Tomato [SMU, Major_Soil, Drainage, Depth, Slope, Temperature, Elevation, PH_Water, PH_Kcl, EC, Na, K, Ca, Mg, CEC, BS, TN, OC, OM, C_N, P2O5, CaCO3, ESP, Ca_Mg, K_Mg, Ca_Mg_k, SUM]

Along with that, to deal with imbalanced data issues of the target attribute, from oversampling techniques Synthetic Minority Oversampling Technique (SMOTE) and Random Over Sampler objects were used to fit the training set.

First using the Maize crop dataset, a prediction experiment was made for ten ML algorithms (Decision Tree, Random Forest, Extra Trees, MLP Classifier, K-Neighbors, Logistic Regression, Gaussian NB, SVM, Xgboost, and AdaBoost). Then among these, the algorithms which return the highest evaluation scores were used to test for the remaining crops. The models' performance was evaluated using the K-fold Cross-validation and Training test 70:30 splitting technique.

From the experimented ML algorithms and settings on Python, the best-suited algorithms for the suitability classification model were identified through the implementation of the Oversampling technique to minority classes.

Table 4: Python Crop-based Algorithms' Prediction Accuracy and Performance Summary

Algorithm	*Crop*	*Sampling Strategy*	*Accuracy*	*Cross-validation*
Decision Tree	Maize	SMOTE "auto"	97.38%	94.31%
	Onion	SMOTE "auto"	99.14%	96.77%
	Rice	SMOTE "auto"	97.06%	95.62%
	Sesame	SMOTE "auto"	98.18%	94.88%
	Sugarcane	SMOTE "auto"	100.00%	98.03%
	Tomato	Random Oversampling "all"	99.59%	98.60%
Random Forest	Maize	Random Oversampling "all"	97.28%	94.81%
	Onion	SMOTE "auto"	98.71%	97.07%
	Rice	SMOTE "auto"	96.88%	95.26%
	Sesame	SMOTE "auto"	97.87%	94.38%
	Sugarcane	Random Oversampling "all"	100.00%	98.28%
	Tomato	SMOTE "auto"	99.82%	98.31%
Extra Tree	Maize	Random Oversampling "all"	97.26%	94.31%
	Onion	Random Oversampling "all"	98.46%	96.59%
	Rice	Random Oversampling "all"	96.89%	94.91%
	Sesame	SMOTE "auto"	98.18%	94.62%
	Sugarcane	Random Oversampling "all"	100.00%	97.18%
	Tomato	Random Oversampling "all"	99.82%	97.72%
Multi-layer Perceptron	Maize	SMOTE 'auto'	96.72%	93.05%
	Onion	SMOTE 'auto'	99.23%	96.63%
	Rice	SMOTE 'auto'	96.48%	94.86%
	Sesame	Random Oversampling "all"	97.46%	94.26%
	Sugarcane	Random Oversampling "all"	99.75%	97.00%
	Tomato	Random Oversampling "all"	99.72%	97.02%
K-Nearest Neighbor	Maize	SMOTE 'auto'	96.45%	92.88%
	Onion	Random Oversampling "all"	96.46%	96.33%
	Rice	Random Oversampling "all"	95.27%	94.37%
	Sesame	SMOTE "auto"	96.12%	95.14%
	Sugarcane	Random Oversampling "all"	99.78%	91.80%
	Tomato	SMOTE "auto"	99.89%	97.54%

5. The Prototype

Initially, the SCSAM deployment plan was to deploy on the Flask web framework. Accordingly, using this popular Python web framework, the Soil-Crop Suitability Prediction Model for Maize crop was developed and tested on the flask framework. However, the bulk sample prediction for the entire soil sample could not be achieved through the framework. It did not allow the user to upload multiple row datasets to the classification model and get the prediction results.

Hence, instead of using flask as the main SCSAM deployment platform, we used a Python GUI library called Tkinter. The deployment process and the Soil suitability Prediction prototype structure is depicted in Figure 1.

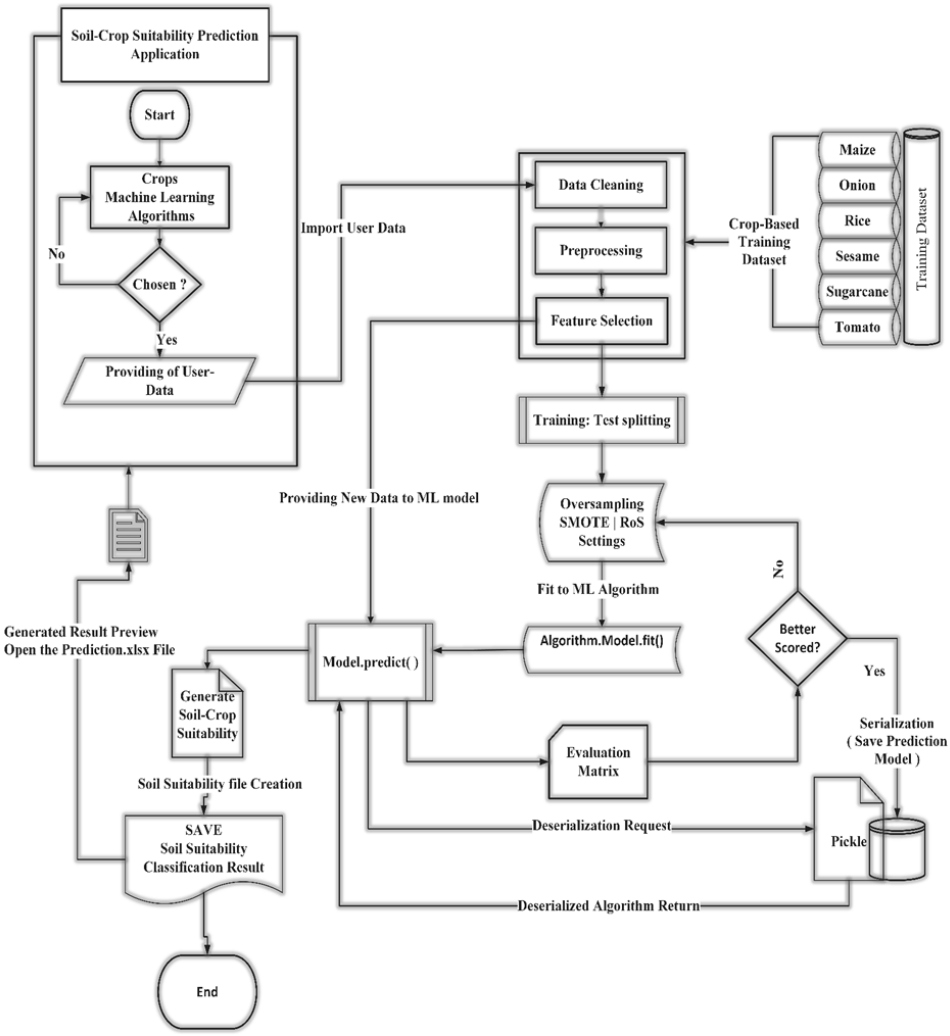


Figure 1: Flowchart of GUI Based Soil-Crop Suitability Analysis Model for Selected Crops

The SCSAM application interface is created with Tk() top-level window (root) object. Figure 2 depicts the soil-crop suitability prediction model prototype user interface.

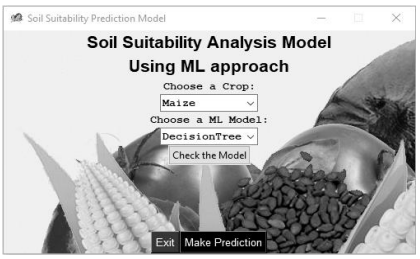


Figure 2: Tkinter Base Soil-Suitability Analysis Model for Selected Crops

- When the SCSAM application starts, it provides an interface to choose the type of crop and prediction algorithm.
- If the user wants to be sure which crops and models are chosen, the "Check the model" button provides that information.
- The "Make prediction" button is used to start the prediction process based on the selected crop and ML algorithm.
- Then it provides a mechanism for users to feed the soil-related dataset into the system.
- To make the user data suitable for ML algorithms, data cleaning and preprocessing works will be done and transform the user data into the prediction process.
- Then in the ultimate stage, the selected ML model makes the suitability prediction and returns the predicted result.

6. Discussion

Model building and prototyping is an approach used to evaluate the solution's functionality, interaction, feasibility, and usability of the product. Once the prototype is developed, it needs to be tested by users.

To test the system's functionality with the user dataset, the prototype was given to senior soil experts, agronomists, and environmentalists in OWWDSE. Twelve user-test evaluation questions were prepared according to [19, 20] and given to them.

In the prototype functionality test, experts tested SCSAM's prediction accuracy with the soil-related data for Maize, Sugarcane, and Tomato crops.

Since the model learning process is initially split from the prediction process, the soil experts highly admired the quick output generation capability for the entire sample. Experts evaluated the prediction performance providing 347 rows of soil dataset. When the prototype returns the prediction result, they compared the output with manually classified crop suitability data. Table 5 shows the total number of correctly classified instances among 347 user-provided data in each crop and the accuracy.

Table 5: The Matching Result of the 347 Rows Containing User Data

Algorithm	*Crop*	*Correctly classified instances*	*Accuracy*
Decision Tree	Maize	320	92.22%
	Onion	334	96.25%
	Rice	332	95.68%
	Sesame	331	95.39%
	Sugarcane	337	97.12%
	Tomato	338	97.41%
Random Forest	Maize	327	94.24%
	Onion	336	96.83%
	Rice	331	95.39%
	Sesame	325	93.66%
	Sugarcane	338	97.41%
	Tomato	336	96.83%
Extra Tree	Maize	324	93.37%
	Onion	333	95.97%
	Rice	328	94.52%
	Sesame	320	92.22%
	Sugarcane	335	96.54%
	Tomato	333	95.97%
Multi-layer Perceptron	Maize	310	89.34%
	Onion	329	94.81%
	Rice	320	92.22%
	Sesame	314	90.49%
	Sugarcane	332	95.68%
	Tomato	332	95.68%
K-nearest neighbor	Maize	309	89.05%
	Onion	331	95.39%
	Rice	315	90.78%
	Sesame	324	93.37%
	Sugarcane	336	96.83%
	Tomato	334	96.25%

Besides, the prototype's performance evaluation questionaries were also supplied to the four senior soil experts and six agronomists found in the enterprise. The questions prepared were designed to obtain the user opinion evaluating the prototype's usability, ease of learning, usefulness, capacity, performance, and output correctness.

The experts ranked that the prototype is lightweight, user-friendly, easy to use, fast and interactive. After evaluating the prototype when we communicate with those professionals, they mentioned that most of the incorrectly classified values in the dataset are tolerable. Even in reality, sometimes, the physical and environmental characteristics of the soil units might have multiple suitability class features and in this kind of situation, experts may decide following their judgments.

7. Conclusion and Future Work

The raw data for this research work was collected from OWWDSE and using diverse machine learning and data mining tools, the collected data were preprocessed. Subsequently, the processed data were used to train the ML algorithms. In the process, various experiments were done on RapidMiner, WEKA, and Python ML tools. From the experiments, the best-suited algorithms and settings were identified in each ML tool. Python and its ML libraries were found to perform better than the others.

Thus, using the best-scored algorithms and settings, SCSAM was developed on the Tkinter platform.

To test prediction capability, the prototype was given to professionals along with 12 evaluation questions. The professionals' evaluation summary and feedback show that the prediction result of the prototype was 88%-97%.

Because of data quality issues, this work could not be able to include all-climate zone data in the ML model. Hence, to make the prototype effective for any large-scale project, future work is required to include the soil-related data of all climate zones. Also, instead of combining both irrigation and rain-fed agriculture data into a single dataset, preparing a separate dataset based on essentials is a future work.

References

[1] OWWDSE, "Hidha-Sombo Small Scale Irrigation Project, Soil and Land Evaluation," Oromia Water Works Design and Supervision Enterprise, Addis Ababa/Finfinne, March 2019.

[2] Hellkamp, A. S., Bay, J. M., Campbell, C. L., Easterling, K. N., Fiscus, D. A., Hess, G. R., McQuaid, B. F., Munster, M. J., Olson, G. L., Peck, S. L., Shaver, S. R., Sidik, and Tooley, M. B., "Assessment of the Condition of Agricultural Lands in Six Mid-Atlantic States," Journal of Environmental Quality, pp. 795-804, 2000.

[3] FAO, "A Framework for Land Evaluation", FAO Soils Bulletin 32, 2nd ed., Vol. 32, Rome, 1976.

[4] FAO, "Soil and Plant Testing and Analysis," in FAO BullETIN 38/1, Rome 13-17, June 1977.

[5] Thom, William O., and Sikora, Frank J, "Soil Testing for Agronomic and Environmental Uses," Vol. 21, No. 4, 2000.

[6] S. Stanford, "The Best Public Datasets for Machine Learning and Data Science," Machine Learning Memoirs Inc.

[7] Oromia Water Works Design and Supervision Enterprise, "Integrated Land Use Plan Study projects," 2001-2019.

[8] M. Mokarram, S. Hamzeh, F. Aminzadeh, and A. Rassoul Zarei, "Using Machine Learning for Land Suitability Classification," West African Journal of Applied Ecology, Vol. 23, No. 1, pp. 63–73, 2015.

[9] Kennedy Senagi, Nicolas Jouandeau, and Peter Kamoni, "Using parallel random forest classifier in predicting land suitability for the crop," Journal of Agricultural Informatics, Vol. 8, 2017.

[10] M. Paul, S. K. Vishwakarma, and A. Verma, "Analysis of Soil Behaviour and Prediction of Crop Yield Using Data Mining Approach," International Conference on Computational Intelligence and Communication Networks (CICN), pp. 766-771, Jabalpur, 2015.

[11] Ratnmala Bhimanpallewar and M. R. Narasinagrao, "A Machine Learning Approach to Assess Crop Specific Suitability for Small/Marginal Scale Croplands," International Journal of Applied Engineering Research, Vol. 12, pp. 13966-13973, 2017.

[12] Ken Peffers, Tuure Tuunanen, Marcus A. Rothenberger, and Samir Chatterjee, "A Design Science Research Methodology for Information Systems Research," Journal of Management Information Systems, Vol. 24, No. 3, pp. 45-77, December 2007.

[13] Vorhies, William, "CRISP-DM – a Standard Methodology to Ensure a Good Outcome," July 2016.

[14] David Camilo Corrales, Emmanuel Lasso, Agapito Ledezma, and Juan Carlos Corrales, "Feature selection for classification tasks: Expert knowledge or traditional methods?", Journal of Intelligent & Fuzzy Systems, Vol. 34, pp. 2825–2835, 2018.

[15] K. J. Anesthesiol, "The prevention and handling of the missing data", Vol. 64(5), May 2013.

[16] Yeshanew Ayalew, "Design and Development of Predictive Model for Potential ATM Card User: Case of Dashen Bank S.C", Unpublished Masters Thesis, January 2019.

[17] Saurav Kaushik, "Introduction to Feature Selection Methods with an Example," Analytics Vidhya, December, 2016.

[18] M. Pathak, "Handling Categorical Data in Python", Datacamp, January, 2020.

[19] Dhandre, Pravin, "Selecting Statistical-based Features in Machine Learning application," Packt hub, https://hub.packtpub.com/selecting-statistical-based-features-in-machine-learning-application/.

[20] F. D. Davis, "Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology," Vol. 13, pp. 319-340, September 1989.

Dictionary Based Word Sense Disambiguation for Amharic

Gashaw Sisay
HiLCoE, Computer Science Programme, Ethiopia
gsisay@gmail.com

Solomon Teferra
School of Information Science, Addis Abeba University, Ethiopia
solomon_teferra_7@yahoo.com

Abstract

A word carrying multiple senses is one source of ambiguity in natural languages that people have to deal with by asking for clarification, using their real world experience or the context as a means to deduce the intended sense of the ambiguous word in that specific communication.

However, computers lack the capacity to identify the intended sense of an ambiguous word that humans have. This paper tries to build a system that can do Word Sense Disambiguation (WSD) on a given sentence by identifying the ambiguous word and using Machine Readable Dictionary (MRD) to infer the most likely meaning of the ambiguous word and conduct performance comparison of the MRD based Lesk Algorithm variants to identify which algorithm is more effective.

The result of the evaluation showed that Simplified Lesk algorithm with stemming has the highest precision score of 0.418 followed by the Original Lesk algorithm with stemming achiving a 0.291 score of precision. Both the Simplified Lesk and Original Lesk algorithm versions without stemming scored a very low precision score of 0.018.

Keywords: Word Sense Disambiguation; Machine Readable Dictionary; Lesk Algorithm Variants; Knowledge Based Approach

1. Introduction

It is a fact that words in natural language have multiple senses (meanings) that can be identified based on context or real world experience. This polysemic (a word having two or more different but related meanings) and homonymic property of words is one source of ambiguity in communication. Amharic also has this feature of natural languages that is a challenge for computers to deal with in processing natural languages. For example, the word *'ሰራተኛ* could mean 'a person engaged in a line of work' or 'a person diligent at his/her work'. Both of these meanings of the word *'ሰራተኛ* are related with the word *'ሰራ* ('work'). Two or more words are considered homonyms if they are either homophones (sound the same) or homographs (have same spelling) or if they are both homophones and homographs but do not have related meanings. For example, the words *'ሠማ* ('to become hot or warm') and *'ሰማ* ('to listen') are homonyms that are homophones but not homographs, the geminating word *'ገና* ('christmass') and *'ገና* ('yet to be') are also homonyms that are homographs but not homophones, the word *'ሳለ* (meanings 'to make a painting' or 'to sharpen' or 'to cough') also show the property of homonymy in that it is homophone and homograph but with completely different meanings [12].

When people face ambiguity in word meanings, they do the disambiguation task from their real world experiences or by asking for clarification, in the case of conversations, which computer systems are incapable of doing. This problem of WSD is an ongoing research area in Natural Language Processing (NLP) where there is a need to come up with better techniques that enable computer systems do disambiguation tasks by themselves.

WSD is the automated activity of finding the meaning of words in a given context when ambiguities arise [1]. NLP activities that require identifying the exact meaning of ambiguous words like Machine Translation (MT), Speech Recognition (SR), Information Retrieval (IR), and Information Extraction (IE) [2] need to incorporate a Word Sense Disambiguation component to increase their accuracy or effectiveness.

One can find NLP research works on Ethiopian languages that cover most of the research areas of NLP even though there is still a lot left to be explored considering the number of languages Ethiopia have and the available features in each language. The research works in [4] (using WordNet), in [5, 9] (using Semantic Similarity) are based on knowledge-based approach, in [3] (Semi-Supervised learning) and [8] are based on machine learning approach for the Amharic WSD problem.

Machine Readable Dictionary (MRD) based approach is a Knowledge Based approach where the disambiguation process is done based on some kind of written document such as human language dictionary, WordNet or thesaurus. This approach uses a list of lexicons and their corresponding meanings in a dictionary to do the WSD [6]. For Amharic WSD problem this approach is unexplored that needs investigation.

The works in [3, 8] that are based on the machine learning approach have performance limited to the ambiguous words they are trained on and preparing a corpus at least for the majority of the ambiguous words is very challenging. The researches that follow the knowledge based approach donot have the problem of preparing training data but they face their own problem of not having an adequate knowledge source and in the worst case none at all. Amharic has a dictionary knowledge source, where WordNet or Thesaurus is not available for NLP applications. Due to this, the work in [4] conducted for Amharic WSD based on WordNet has to construct Amharic WordNet limited to 10,000 synsets and 2,000 words from a dictionary.

Creating a practical solution for the Amharic WSD problem based on these researches has its own challenges. It is impractical to prepare a training data for all ambiguous words which makes machine learning researches unsuitable and the WordNet based Amharic WSD has its own limitation of word size (having only 2,000 words). On the other hand, using dictionary as a knowledge source, one can create a practical solution that only requires having a database of Amharic-to-Amharic dictionary. This paper aims at creating a solution for the Amharic WSD problem using Amharic-to-Amharic dictionary of more than 25,000 words [7] to infer the most likely meaning of the ambiguous word and provide performance comparison of MRD based Lesk Algorithm variants.

2. Related Work

The work in [3] was conducted to develop Word Sense Disambiguation Model for Amharic Words using Semi-Supervised Learning approach. The experiment was conducted using five selected ambiguous Amharic words with a total of 1031 datasets, two clustering algorithms (simple K-means and EM) and five classification algorithms (AdaboostM1, bagging, ADtree, SMO and Naïve Baye). The experiment result shows performance score of 88.47% using ADtree, 87.40% using SMO and 83.94% using AdaboostM1 and window size of 2-2 and 3-3 can be an optimal window size for Amharic WSD systems development using Semi-Supervised Learning approach.

The research work of Solomon Mekonnen [8] is machine learning based on the Amharic WSD problem by applying a supervised machine learning approach to a corpus of Amharic sentences. In this study, a total of 1045 English sense examples for the five ambiguous words such as አጠና (eTena), መሳል (mesal), መሣሣት (me`sa`sat), መጥራት (metrat), and ቀረጸ (qereSe,) were used. Datasets were collected from British National Corpus (BNC) and the sense examples were translated to Amharic using dictionary and preprocessed to make them ready for the experiment. Then, Naive Bayes classifier was employed from Weka 3.62 package to perform the supervised learning on the preprocessed dataset using 10-fold cross-validation. The author evaluated the classifiers for the five ambiguous words and achieved accuracy within the range of 70% to 83% which was very encouraging and concluded that window size of 3-3 on both side of the ambiguous words is enough to disambiguate.

The effectiveness of semantic similarity measure in WSD for query expansion was investigated in [9]. A simple Word-Net (words with their synonym and gloss definitions) was constructed to be used as a knowledge base in identifying the sense of a given query by applying semantic similarity measure.

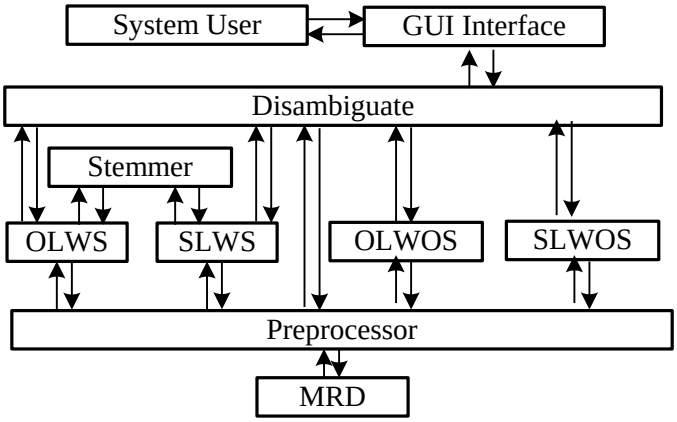
Using the idea of Lesk algorithm, WSD was performed with two methods: gloss to gloss and synset to gloss by comparing information associated with its synonyms and gloss definition with reference to Amharic Word-Net. The combination of the two disambiguation methods formulates the modified query and used for expanding the original query. Finally, the query expansion module is integrated with Information Retrieval system to show the enhancement of Amharic IR system performance. Accordingly, the study has reported that the method using synset for query expansion registered performance of 59% F-measure. Besides, this method registered 6% improvement from the original query.

A WordNet based work by Seid Hassen Yesuf and Yaregal Assabie [4] is the only Knowledge-Based research in the Amharic WSD problem. This research work developed a word sense disambiguation system for Amharic using Amharic WordNet. Due to the nonexistence of WordNet for Amharic, the authors constructed a WordNet containing 10,000 synsets and 2,000 words using Amharic dictionary as a source. The use of WordNet provided senses of words along with the relation these words have with other words to be used in the disambiguation process. Evaluation of the system was made using precision and recall by taking into consideration context window sizes and morphological analyzer effect on 200 randomly selected sentences containing ambiguous words with two or three frequently used senses taken from newspapers and linguistic experts as a test data. The result shows that using this approach an accuracy of 57.5% and 80% can be achieved with and without morphological analyzer, respectively.

The work in [6] is for the Bangla language using a dictionary based approach which is the most related research work to this paper. The research objective was to come up with a system that can disambiguate noun, adjective and verb ambiguous words in Bangla language sentences which have multiple senses by using a dictionary lookup. The research proposed two algorithms where the parsing algorithm creates a parse tree from an input sentence to verify the words existence in the Bangla Language dictionary and the validity of the sentence structure while the detect algorithm looks for ambiguous words in the input Bangla sentences by looking for which word is most overlapped with the dictionary definitions and also its actual meaning. Evaluating the system with limited sentences and without considering semantic features of a sentence, an overall accuracy of 82.40% was achieved.

3. The Proposed Solution

The WSD system uses an Amharic sentence/phrase written in Ethiopic script (fidel) and entered into the system by the user as input. The second input to the system is the dictionary definitions of a word fetched from the MRD when requested by the WSD system. Each of these two inputs go through processes (preprocessing, stemming and algorithmic evaluation) to detect and disambiguate ambiguous words. After this, the most likely dictionary definition of the detected ambiguous word is returned as an output. Figure 1 shows the design of the WSD system which is followed by each module's description.



OLWS - Original Lesk With Stemming OLWOS - Original Lesk With Out Stemming
SLWS - Simplified Lesk With Stemming SLWOS - Simplified Lesk With Out Stemming

Figure 1: Design of the WSD System

A. GUI Module

The user interface has three parts. The first one is where users input a sentence in Ethiopic script to be disambiguated and a button to start the process. The second part shows the result of the disambiguation process in short while the last part shows in detail each step of the disambiguation process. Figure 2 shows the graphical user interface.

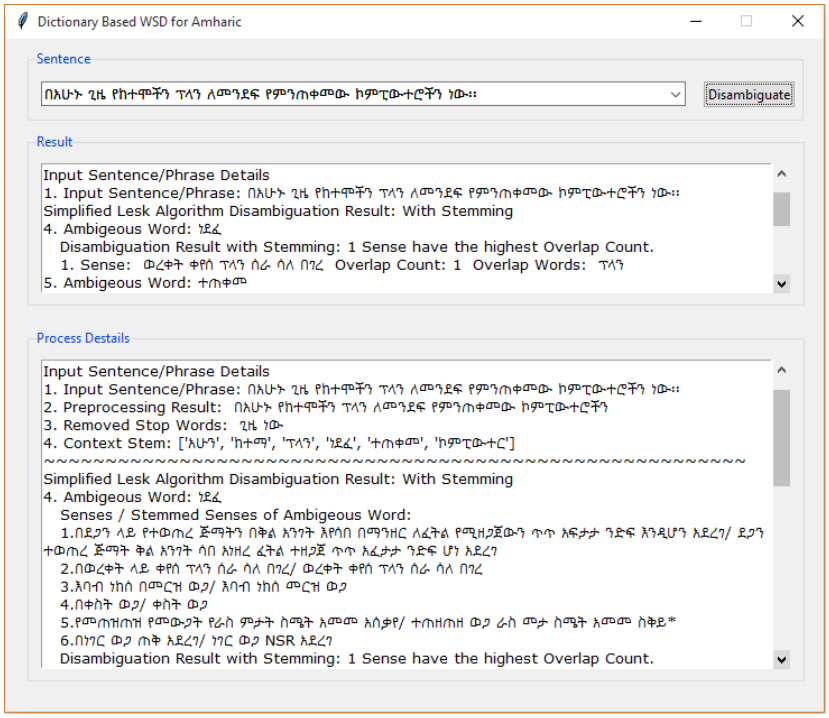


Figure 2: GUI

B. Preprocessing Module

This module removes different components to make the sentence/phrase more suitable for the disambiguation task by breaking down the input sentence/phrase into a list of words (tokenization) and removing items that are not relevant for the disambiguation task like punctuation marks, numbers and other symbols that are not letters. Stop words (words that do not carry meaning) are also removed using a pre-prepared list of stop words as a reference from the MRD. This module is used by the disambiguate module to preprocess user input sentence/phrase while original lesk with stemming, original lesk without stemming, simplified lesk with stemming and simplified lesk without stemming modules use it to preprocess dictionary definitions fetched from the MRD.

C. Disambiguate Module

This module handles the coordination activity of all the other modules of the WSD system.

D. Stemmer Module

This module is constructed by integrating the HornMorpho L3 library (a morphological analysis tool for Amharic, Afan Oromo and Tigrigna languages) with a code to extract the stem/root part of the morphological analysis result. The module stems a single word passed to it as a parameter and it returns the stem of the word. If the word does not have a stem for some reason, the module returns the phrase 'NSR' (No Stemming Result). To stem a word, it is saved in a text file with UTF-8 encoding so that the file content is considered as a Unicode text. L3.anal() method of the HornMorpho library with parameters of language marker 'am', input text file, UTF-8 encoded text file for output, citation enabled, grammar and raw result disabled is run. The UTF-8 encoded output text file is then processed to extract the stem/root of the word among other results of the morphological analysis.

E. MRD Database Module

This module organizes the words with their definitions and a pre-compiled list of stop words using five tables in a single database. *StopWordList* table holds only Amharic stop words compiled from [10, 11] for the purpose of referencing stop words so that the system can detect and remove stop words both from the input sentences and dictionary definitions of words. *WordList* table holds a list of words with their POS and the number of definitions it has compiled from the Amharic-to-Amharic dictionary [7]. *DefinitionList* table contains the definitions of all the words from the *WordList* table with their definition number. *ReferenceList* table holds one word referring to another word where the first word does not have a definition entry in the dictionary while the second word has. *SubWordList* holds list of words that are derived from other words to show which word is the root and which one is the derived word. Figure 3 shows the structure of the MRD Database.

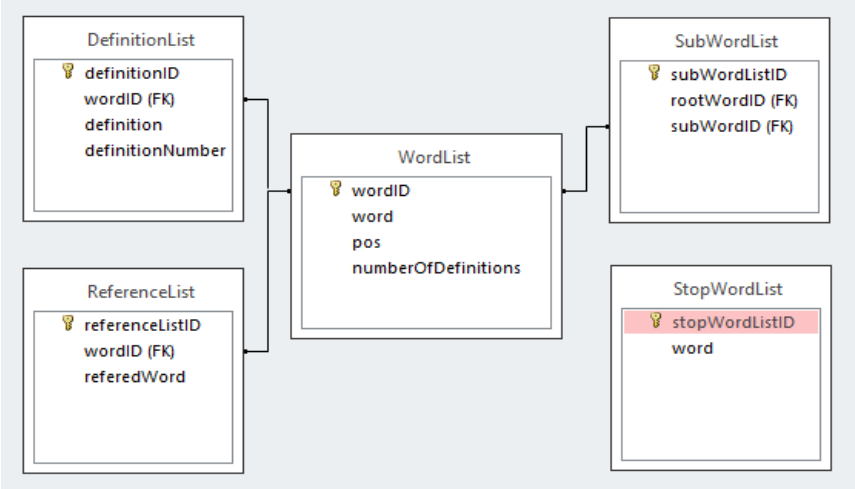


Figure 3: Structure of the MRD Database

F. Original Lesk Algorithm with Stemming (OLWS) Module

After applying preprocessing and stemming on the input sentence/phrase, this module looks for ambiguous words in the input by checking how many definitions each word has in the *WordList* table. A value of two or more number of definitions qualifies a word to be ambiguous. For each ambiguous word detected, dictionary definitions are fetched, preprocessed and stemmed to do count of common words among the definitions of each word in the input and the senses of the ambiguous word. The ambiguous word definition with the highest overlap count is returned to the Disambiguate module. 'Zero or Equal Overlap Count' is returned for zero or equal number of overlap.

G. Original Lesk Algorithm without Stemming (OLWOS) Module

It is similar in function and operation to the previous module except that all the overlap

calculations are done without applying stemming to the words involved in the process.

H. Simplified Lesk Algorithm with Stemming (SLWS) Module

This module preprocesses and stems the input and then looks for ambiguous words in the input sentence/phrase by checking how many definitions each word has in the *WordList* table. A value of two or more number of definitions qualifies a word to be ambiguous. For each ambiguous word detected, dictionary definitions are fetched, preprocessed and stemmed to do count of common words among the definitions of each word in the input and the context. The ambiguous word definition with the highest overlap count is returned to the Disambiguate module. 'Zero or Equal Overlap Count' is returned for zero or equal number of overlap.

I. Simplified Lesk Algorithm without Stemming (SLWOS) Module

This module is similar in function and operation to the previous module except that all the overlap calculations are done without applying stemming to the words involved in the process.

J. Flow of Operation

Disambiguation of the test data follows a sequence of operations coordinated by the Disambiguate module. The list provided below shows the steps taken.

i. User inputs a sentence to be disambiguated and clicks on the Disambiguate button. The Disambiguate module is called and the disambiguation process starts.

ii. Disambiguation module gives control with the input sentence/phrase to Simplified Lesk with stemming module. This module disambiguates the input after applying preprocessing and stemming it and returns the result with control to the Disambiguate module. After this, Disambiguate module passes the result to the GUI module to be displayed.

iii. Disambiguation module gives control with the input sentence/phrase to Simplified Lesk without stemming module. This module disambiguates the input after preprocessing it and returns the result with control to the Disambiguate module. After this, Disambiguate module passes the result to the GUI module to be displayed.

iv. Disambiguation module gives control with the input sentence/phrase to Original Lesk with stemming module. This module disambiguates the input after preprocessing and stemming it and returns the result with control to the Disambiguate module. After this, Disambiguate module passes the result to the GUI module to be displayed.

v. Disambiguation module gives control with the input sentence/phrase to Original Lesk without stemming module. The module disambiguates the input after preprocessing it and returns the result with control to the Disambiguate module. After this, Disambiguate module passes the result to the GUI module to be displayed.

vi. Disambiguation module exits and control is returned to the main loop of the system.

4. Discussion

The performance of the algorithms used in the disambiguation process was measured in terms of precision and recall. Precision was defined as the *number of words correctly disambiguated over total number of words disambiguated* while Recall was defined as the *number of words disambiguated over the total number of words attempted to be disambiguated*, according to the definition of precision and recall measure for WSD proposed in [13].

Test data (a set of sentences for each and every sense of 25 ambiguous words selected from [7] with two or more distinict senses) and MRD (a database of words with their senses acquired from [7] and a list of stopwords compiled from [10, 11]) were given to the WSD system as input and the result was analysed.

MRD based WSD depends on the count value of common words between context (the dictionary definition of the words in the context) and dictionary definitions of the ambiguous word. Since this count value is based on comparing words similarity, words that are derived from the same root word are

considered different words and fail the common word count. To remedy this situation, stemming was applied on every word before any word similarity comparison was made in the algorithms with stemming version. Table 1 shows the stemming result of the test data and definitions/senses from the MRD.

Table 1: Stemming Result

Data Source	*Correct*	*Incorrect*
Test Data	94%	6%
Definitions/Senses	95%	5%

Application of stemming in the Original Lesk Algorithm resulted in 0.27 and 0.84 increase in precision and recall values, respectively. In the Simplified Lesk Algorithm, it has also caused 0.4 and 0.82 increase in precision and recall values, respectively.

Evaluation of the algorithms' performance (Table 2) shows that Simplified Lesk with stemming has the highest precision (0.42) while Original Lesk with stemming produced the second highest precision (0.29). Out of all the disambiguation attempts made, Simplified Lesk with stemming resulted in a small error of 5% whereas 25% error had occurred in the Original Lesk with stemming. The lowest result for precision (0.02) is recorded in both the Original Lesk and Simplified Lesk algorithms without applying stemming.

Table 2: Performance in terms of Precision and Recall

Algorithm	*Precision*	*Recall*
Original Lesk with Stemming	0.29	0.96
Simplified Lesk with Stemming	0.42	0.96
Original Lesk without Stemming	0.02	0.13
Simplified Lesk without Stemming	0.02	0.15

5. Conclusion

This paper presented a system for the Amharic WSD problem originating from Polysemy and Homonymy using Amharic-to-Amharic dictionary as a knowledge source and provided performance comparison of the MRD based Lesk algorithm variants used in the disambiguation process.

In conclusion, stemming has a positive influence in precision and recall by converting words into their root form for detection and disambiguation. However, it has also caused the number of incorrect disambiguation results to increase as more words are disambiguated than before. In both with and without stemmer versions of the algorithms, Simplified Lesk with stemming has the highest performance score and the application of stemming enhanced the algorithms' performance.

References

[1] Roberto Navigli, "Word Sense Disambiguation: A Survey", ACM Computing Surveys, Vol. 41, No. 2, February 2009.

[2] Alok Ranjan Pal and Diganta Saha, "Word Sense Disambiguation: A Survey", International Journal of Control Theory & Computer Modeling (IJCTCM), Vol. 5, No. 3, July 2015.

[3] Getahun Wassie, Ramesh Babu, Solomon Teferra, and Million Meshesha, "A Word Sense Disambiguation Model for Amharic Words using Semi-Supervised Learning Paradigm", STAR Journal, July-Sep 2014: 3(3), pp. 147-155.

[4] Seid Hassen Yesuf and Yaregal Assabie, "Amharic Word Sense Disambiguation Using WordNet", 5th International Conference on the Advancement of Science & Technology (ICAST-2017), Bahir Dar University, Ethiopia, May 2017.

[5] Teshome Kassie, "Word Sense Disambiguation for Amharic Text Retrieval: A Case Study for Legal Documents", Unpublished Masters Thesis, Addis Ababa University, 2009.

[6] A. Haque and M. M. Haque, "Bangla Word Sense Disambiguation System Using Dictionary Based Approach", Proceedings of the International Conference on Advances in Information & Communication Technology, 2016.

[7] አማርኛ መዝገበ ቃላት፤ የኢትዮጵያ ቋንቋዎች ጥናትና ምርምር ማእከል፤ አዲስ አበባ ዩኒቨርሲቲ፤ የካቲት 1993.

[8] Solomon Mekonnen, "Word Sense Disambiguation for Amharic Text: A Machine Learning Approach", Unpublished Masters Thesis, Addis Ababa University, 2010.

[9] Samrawit Zewdneh, "Word Sense Disambiguation using Semantic Similarity for

Query Expansion in Amharic Information Retrieval", Unpublished Masters Thesis, Addis Ababa University, 2014.

[10] Tihitina Petros Degsew, "Modeling and Designing Amharic Query System to Bilingual (English-Amharic) Databases", Unpublished Masters Thesis, Addis Ababa University, 2014.

[11] Mequannint Munye Zeru, "Amharic-English Bilingual Search Engine", Unpublished Masters Thesis, Addis Ababa University, 2010.

[12] PEDIAA, www.pediaa.com/difference-between-polysemy-and-homonymy/, last accessed on April 2019.

[13] Rada Mihalcea and Dan Moldovan, “A Highly Accurate Bootstrapping Algorithm for Word Sense Disambiguation”, International Journal on Artificial Intelligence Tools, World Scientific Publishing Company, Vol. 10, Nos. 1 & 2, 2001.

Security-Integrated Enterprise Architecture Framework (SIEAF)

Heven Hailu
HiLCoE, Software Engineering Programme, Ethiopia
h.hailumst@gmail.com

Dereje Teferi
HiLCoE, Ethiopia
School of Information Science, Addis Ababa University, Ethiopia
dereje.teferi@gmail.com

Abstract

Sophisticated and new forms of threats to information assets makes information security one of the major concerns of enterprises. It has been argued that the integration of security in an EA (Enterprise Architecture) strengthens the effectiveness and efficiency of enterprises. Moreover, causes of system attacks are associated with the architecture of systems. Therefore, designing security oriented EA is indispensable. On the other hand, EA is an approach that can serve as a means to govern Systems of Systems (SoSs) and to study security from the perspective of all elements that influence the enterprise's system architecture. In this paper, various security and EA frameworks and their integration with each other is explored for designing SIEAF. The use of Zachman for modeling the high level of EA and Sheerwood Applied Business Security Architecture (SABSA) for security and the relation of both is described step by step. TOGAF-ADM (The Open Group Architecture Framework-Architecture Development Method) is recommended for the EA development process at mid-level following the Zachman framework guidance. At the same time, SADM (Security Architecture Development Method) should be integrated with TOGAF-ADM to address information security. Ultimately, ArchiMate, MAD (Mal-activity Diagram) and/or ISSRM-DM (Information System Security Risk Management-Domain Model) should be mapped with each other at the implementation level.

This framework comprises of the EA as well as information security risks and control mechanisms on all levels of the EA. In a nutshell, the defined framework is capable of measuring the security of the EA as a whole and provides security countermeasures based on the information flow at each layer.

Design science research method was used in this study. The design of SIEAF was based on the literature on security oriented EA frameworks as well as interviews of 17 security and EA practitioners and experts. As a result, the feedback was positive and hence the framework is considered to be both usable and useful. So the defined framework is capable of modeling all possible attack scenarios or malicious activities and their sequential presentation of operations based on the information flow of EA.

Keywords: Enterprise Architecture; EA Frameworks; Information Security Principles; Security Metrics; Security Artifact; Security Risk Management

1. Introduction

There are various literature describing enterprises consisting of numerous assets and need to be protected from being accessed by unauthorized parties, asset modification and system destruction [1]. Concerning the enterprise systems, describing the type of security that is required by a specific asset in a particular layer of the organization is crucial for securing an enterprise and its emergent behavior [2]. Designing security that assures the balance between information risks and controls is indispensable [3]. However, "based on the increasing requirements of high-quality and reliable systems, developing security-integrated EA is very challenging" [2].

EA is a comprehensive picture of enterprises that mainly aims on integration, service oriented architecture and a means of examining security [1]. More importantly, EA is an approach that manages changes, align IT with business, decrease complexity which greatly reduces costs and manages stakeholders' concerns that helps to do efficient decision making. Therefore, EA can be defined as a method that evidently demonstrates an integrated structure of information, application and IT infrastructure. On the other hand, EA investigates the

emergent features of the interaction of IT and business which is beyond the IT and business architecture or even both side by side [3]. Comparisons of various security and EA frameworks, methods, and modeling languages and their integration is explored in this paper.

There are several discussions and researches concerning the security of enterprises and defined several approaches to address the subject. One of these studies, which currently is accepted by industry standards, is the Gartner Enterprise Information Security Architecture (GEISA). The focus of this enterprise security architecture is the description of the structure by considering its compatibility with EA programs and the collaboration of both [3]. However, it has no specific methodology and modeling language for implementation and hence, modeled information security in a separate system which could lead to a potential new vulnerability. Further works in this area of study include the Zachman and SABSA frameworks. The Zachman framework is adapted to include information security [3]. But it does not specifically address security aspects. Besides, as it is not a method rather a framework, it doesn't include a specific methodology for implementing both security and EA. The SABSA framework is "the most outstanding attempt of a holistic Enterprise Information Security Architecture (EISA)" [2]. But it is adapted mainly for information security which could not abstract the entire EA including System of Systems (SoSs) and their emergent behavior equivalent to the Zachman framework. So it can be said that there have not been sufficient and comprehensive approaches that consider security integration from an early phase of EA development that include elements of any information system design which are frameworks, methods and modeling languages and information system security domain model. They depend on predefined vulnerabilities, as security is designed implicitly in the EA which could not be extracted and leads to lack of analyzability of the security features.

EA framework has an important capability in optimizing business. However, a few EA frameworks provide a comprehensive security integrated enterprise architecture framework. The aim of this paper is to develop quantifiable security metrics and its countermeasures in a graphical modeling language which comprises security and EA components and their integration with each other. This framework contains both the EA as well as the security risks and measurements on all levels of EA. It defines security baseline, gap analysis, metrics, compliance, policy, standards, guidelines, and procedures.

2. Related Work

2.1 Enterprise Architecture Frameworks

As the demand of managing complex enterprise systems has been increasing, the role of EA is taken as important to facilitate communication among the organizations' business plans and define complete frameworks [2]. The applicability of EA frameworks to abstract complex systems including its emergent behavior needs to be explored. Emergent behavior is a behavior of systems that does not depend on its individual parts, but on their relationships to one another. "All the Frameworks are applicable to a single system, but not all the frameworks are suitable for SoS" [4].

One of the most popular EA frameworks is DoDAF (Department of Defense Architecture Framework). It is a well-defined architecture and its products and their requirements are detailed and structured on four views which are the *overall view*, *operational view*, *system view* and *technical view*. However, DoDAF depends on "interface" which is not applicable to the complex systems with emergent behavior due to the fact that missing of a system, subsystem or component will make the integration impossible [4]. So DoDAF could not be applicable as a framework guide in the EA development.

MODAF (Ministry of Defense Architecture Framework) is defined to structure the integration of systems in an organization. It is the extension of DoDAF to support capability management process which added two views, strategic and acquisition [2]. But it is not suitable for the development of ultra-large systems. This is because it is product-centric and can be implemented in a single system.

QGEAF (The Queensland Government Enterprise Architecture Framework) is mainly designed to scale up the compatibility and decrease the cost of IT

infrastructure. It organizes the enterprise resources including the infrastructure [2]. Nevertheless, as it is specific to a particular system, it would not be suitable to abstract the entire EA of large systems.

TOGAF (The Open Group Architecture Framework) is described mainly to integrate IT with business. Moreover, TOGAF-ADM is being used by users of TOGAF as a guide for planning architecture. This is due to its flexibility and might be used with deliverables from other frameworks such as Zachman. ADM is a generic architecture with features of loose coupling and emergent behavior of complex systems. So TOGAF-ADM is used in the EA development process following the Zachman framework guidance.

FEAF (Federal Enterprise Architecture Framework) is defined mainly to avoid redundancy and improve resource sharing among organizations. It also includes characteristics of the Zachman framework. Though it could be suitable to model the system of systems, it cannot complement with SABSA better than Zachman as both Zachman and SABSA frameworks are alike in their structure and content.

Zachman provides a way for classification of an organization's architecture to organize "primitive" architectural information; no specific models, no methodology and no notation. The Zachman framework can be used to describe any complex entity and it is not product specific. These features of Zachman are advantageous and better than other EA frameworks for developing SoS architecture with emergent behavior since the architecture developer will have the freedom to incorporate the latest modeling techniques of emergent behavior. "The SoS architecture is a layered architecture model which allows different developers to work in parallel and ensures that changes in one layer of the protocol do not interfere with operations above and below that layer" [4].

Generally, most of the frameworks are not suitable to develop ultra large systems by including emergent behaviors [4]. Besides, all the current EA frameworks have the same aim, which is efficient use of IT to support business needs that decrease complexity and share organizational resources. However, none of these EA frameworks support an approach which measures security of an enterprise based on the existing EA products [2].

2.2 Information Security Architecture

Based on extensive literature, recent and universally accepted EISA standards are GEISA, Zachman, and SABSA. The Zachman framework is adapted to include security but does not explicitly address the security issues in EA and hence the SABSA framework is intended to fill the gaps not addressed by Zachman. Furthermore, GEISA focuses on description of structures but it does not include a specific method for implementation and it is abstract and theoretical compared to SABSA. SABSA is more practical and includes its own specific method for carrying out requirements engineering. Therefore, SABSA is the most outstanding holistic EISA that can complement with Zachman [3].

SABSA closely follows the work by Zachman, although it has been adapted to security. Moreover, there is no independent security component in FEAF. DoDAF has no any specific viewpoint of security. MODAF considers security architecture implicitly similar to DoDAF that could not be extracted and leads to lack of analyzability of the security aspects in a given organization.

In summary, it can be said that all existing EISA focus mainly on the description of structure that could comply with EA programs. This eventually led to a potential vulnerability and lack of analysability of security variables. However, none of these frameworks provides an approach that enables an enterprise to track all possible malicious activities and their sequential presentation of operations graphically based on the information flow of EA.

2.3 Security-Oriented Enterprise Architecture Frameworks

2.3.1 Enterprise Architecture Security Assessment Framework (EASAF)

According to Alshmmari, "The most efficient approach for quantifying the overall information security of a given enterprise is to measure it with respect to its EA" [2]. The approach used in [2] provides an assessment framework focused only on the middle-level (Method) abstraction that comprises the SADM and the TOGAF-ADM for security and

EA respectively. This approach needs to consider the higher level presentation (the framework) which is used to model the structured set of essential components of security and EA and their relationships; the low-level component that defines a set of rules to interpret the meaning and alignment of security and EA concepts and their relationships; and the security domain model used to understand the concepts of security criterion, vulnerability, event and risk constructs and tracks the scope of security control that avoids visualization difficulties.

This paper provides a technique to quantify possible attack developments based information flow of an enterprise via the four essential components that define the high to low-level presentation of security and EA principles. These are frameworks, methods, domain model and modelling languages.

2.3.2 Designing Secure Enterprise Architecture

Bosch [3] claims that designing secure EA approach could possibly benefit both the security and EA disciplines. These benefits include addressing the security requirements and provision of holistic view where security is adequately addressed from both security and EA perspectives.

SEAF tried to address the problem of security of an organization by assessing the as-is architecture, defines the gap and finally states the to-be security extension on the ArchiMate tool in the implementation level. The security extension in ArchiMate provides the various concepts needed to include security in the architecture specification, specifically vulnerability, threat, risk, security mechanism, and security policy. These security concepts are included in SABSA to fulfil security that is not in ArchiMate [2]. However, this kind of iterative method has a disadvantage. For instance, if the approach is applied to all the assets of a system, the EA model will blast (i.e., it will be impossible to keep track of these assets, controls and relations). Generally, the security extension on ArchiMate could bring complications of visual presentation of outcome. It is not structured and understandable especially when the method enhances after each application loop [6].

SIEAF provides security domain model at the implementation level to identify important assets of an enterprise and prioritize by aligning with EA modelling language. This avoids visualization difficulties of results as asset modelling and security tracking increases after each application round.

2.3.3 MF4EASP – The Method Framework for Developing Enterprise Architecture Security Principles

The work in [7] defined a framework for integrating information security with EA management, based on the organization's EA security principles instead of only relying on technical security mechanisms provided by the existing systems and software.

This method needs to consider detail and hierarchical EA modelling to determine the expected components from the organizational context. It is highly abstracted and lacks visualization as it doesn't consider security domain model.

So the aim of this paper is to develop a framework which comprises both EA as well as the security risks, measurements and security domain model on all levels of EA hierarchically from high to low levels.

3. The Proposed Security-Integrated Enterprise Architecture Framework (SIEAF)

The main goal of this paper is to define a framework that consists of EA principles, EA artifacts and EA metrics as well as a set of EA security related principles, security artifacts and quantifiable security metrics that measure the enterprise's security level from the perspective of business architecture and data flow at each layer. It also defines assets, risks and countermeasures in graphical model. This framework provides an organization's security experts a possibility to track the entire security change management, specifically the security of an enterprise's elements and define a rule for transforming the predefined security version of EA components. Much of the existing work has addressed security as a separate discipline when organizational architecture is developed [2]. The synergic benefits of security and EA have not been sufficiently captured [3]. Security is not addressed until the strategic design of an enterprise is

established and even business is running [6]. This way of treating enterprise's security has brought various information security problems.

The framework, SIEAF, consists of four components: *Frameworks*, *Methods*, *Domain-Model* and *Modeling Languages*. A *Framework* subdivides the structures of the viewpoints in different ranges including the relationships between these domains. This includes the SABSA and Zachman frameworks for both security and EA respectively. The second component of the framework is the *Method*, which defines a work product that describes the processes for developing the structural descriptions of the main framework step by step. This methodology comprises SADM and TOGAF-ADM for detailed explanation of both security and EA principles, artifacts and metrics. The third component is a *Modeling Language* which is defined by a consistent set of rules that interpret the meaning and alignment of security and EA concepts and their relationships. It consists of MAD and ArchiMate that describe the alignment of elements and present both the security risk treatment module and the EA, respectively. The fourth component is the *Domain-Model* which presents security requirements, security control and especially the security risk treatment element which could not be found in both modeling languages. ISSRM-DM is used to help information security developers understand the concepts of security criterion, vulnerability, event and risk constructs. Mapping between this domain model and the two modeling languages is very significant to identify and prioritize important assets and present possible attacker's action, behavior and its impact on entire assets of an enterprise.

3.1 The Framework: Corresponding Integration of Zachman and SABSA Frameworks

To achieve full-fledged and security-integrated EA framework, both Zachman and SABSA frameworks and their integration have been chosen. This is because according to Bondar *et al.* [4], "the agility is more important than technology skill. The 21st century is about agility, adjustment, adapting and creating new opportunities." Hence, both Zachman and SABSA frameworks are agile and proactive and they could outline more comprehensive structure which is the basis for the design, choice, and implementation of all subsequent EA and security controls. Furthermore, both of them are open standards. They provide comprehensive views and they complement with each other regarding their contents. They also provide the relationships between the concepts in their matrix. Except one layer, all the structures of Zachman are similar with the structures of SABSA. They are used to describe standards, for example, standards for enterprises' information systems. Each cell of the frameworks contains such a series of standards for organization's information systems [8]. Zachman matrices provide views which document the system's functional structure including key functional elements, their responsibilities, the interfaces they expose and the interactions between them and SABSA like Zachman describes the security architecture and its assurance for the enterprise based on the information flow designed by Zachman [3]. Moreover, TOGAF-ADM aligns with the first four rows of the Zachman and SABSA frameworks. For the output of each of the phases, it illustrates which views and view points in both Zachman and SABSA are applicable to the specified phase. While integrating both frameworks, iterative method is used, that is, once one of the layers of Zachman is filled in, the complementing layer in SABSA is also constructed.

Therefore, it can be concluded that the integration of SABSA and Zachman side by side to design a comprehensive approach is a perfect complement.

3.2 The Method (Processes Described to Design the Framework): Complementary Integration of TOGAF-ADM and SADM

TOGAF-ADM is flexible in that it may be used with a set of deliverables from another framework, or it may even be used in conjunction with the Zachman and SABSA frameworks [4]. TOGAF-ADM allows an organization to choose to bypass or tailor any part of the process as required. ADM is a generic method for architecture development, which is designed to deal with most system requirements. This suits well with loose coupling and emergent behavior of a SoS environment.

A widely accepted standard in EA development, especially in prescriptive and step wise processes, is

TOGAF-ADM, designed by the Open Group [9]. It is the de-facto industry standard and provides a structured approach of designing enterprise architectures. Since it is also an open standard, it is well suited for the purpose of the method. However, according to literature no single development method is the industry standard.

The TOGAF-ADM needs to include the important elements that help develop the EA method. Those are EA principles including its design, EA artifacts and EA metrics. EA artifacts are products resulting from the architecture process. Artifacts are created to "describe a system solution, or state of the enterprise" [9]. EA metrics provide understanding and knowledge whether the EA is providing a value to the business or not. Three of these elements will be complemented with the elements of SADM, that is, security principles, security-related EA principles, security-related artifacts and security metrics. According to the Open Group [9], "Within TOGAF, the role of security is acknowledged but not designed and completed". The work of the enterprise architect and security practitioner is often separated while it is really needed to be fully integrated. Hence, a method called SADM that is responsible for developing security architecture is required. It is an iterative process that enables enterprise architects to develop, assess and maintain the security assessment framework [2]. SADM defines security related EA principles because in terms of information security, EA principles that have an impact on security must be identified, assessed and maintained to achieve a more secure architecture and EA principles play a major role in deploying the organization's business strategy and once properly defined, they have a major impact on achieving the organization's vision. TOGAF-ADM and SADM are methods which are used to define the EA principles and information security principles, respectively.

3.3 The Modeling Language (rules to interpret the meaning of concepts): Integration of ArchiMate and ISSRM-DM and MAD

ArchiMate is an Open Group standard that offers an open and independent modeling language for EA. It is widely known and used in different consulting firms and supported by several tool vendors. The ArchiMate language consists of modelling elements that represent real life entities and relationships between them [6]. This model is suitable to align or map with ISSRM-DM and MAD.

ISSRM-DM protects the essential IS components. It could be used to represent asset-related, risk-related and risk treatment-related concepts. The outcome of each gives fundamental understanding about assets, vulnerabilities and threats and defines the scope of the information system security as it comprises both the business and IS assets in general [6]. Possible risks could be mitigated through implementation of controls. Search of these countermeasures is made via risk analysis. Unfortunately, after controls are defined for some particular asset, it is not visible to investigate how implementation of these countermeasures will influence the whole EA. That is why the necessity of integrating the EA and the security risk management (SRM) domain model from the initial stage remains actual and motivating. Moreover, ISSRM-DM is capable of defining the same terminology for two modeling languages, ArchiMate and MAD.

According to [6], MAD presents SRM in more detail and has smoother and continuous move between requirement engineering and design stages. Moreover, it has the biggest emphasis on the attack process and attacker behavior. In other words, it helps to add malicious activity into normal work process. So MAD is used to define the security notions for describing security structure and to graphically present ISSRM risk treatment-related concepts in a "swim lane" which would be implemented in the EA (ArchiMate modeling language). Therefore, ISSRM-DM and the two modeling languages align with each other to graphically present secure EA. The proposed SIEAF is shown in Figure 1.

The proposed framework is presented in the SABSA enterprise security architecture framework using ArchiMate Modeling language. This is because SABSA develops business-driven, threat and opportunity focused enterprise security and information assurance architectures; delivers security infrastructure solutions that traceably support critical business initiatives; and provides a future-proof

framework for information management [11]. The proposed framework is developed at contextual, conceptual, logical, functional, physical, component and operational levels.

The *Contextual Architecture layer* (business view) is the main function to identify and extract security requirements from the business environment. In this layer, EA and information security goals and objectives are presented via Zachman and SABSA, respectively. Information security goals are used to realize the protection of the EA components. Hence, based on the security requirements, the business risk model is set to realize/influence the business process model from an early stage of the EA development process. Besides, this layer addresses employee's knowledge which is presented in the ISSRM-DM as an asset that needs security.

The *Conceptual Architecture layer* (Architects view) generalizes the business concerns and considerations into policies and procedures. EA policies and Security Architecture policies produce information security architecture based on the data flow analysis policies that track the business flow among different high and low-level organizational elements.

The *Logical Architecture layer* (Designer's view) defines detailed security controls and management objectives, for example, access control and monitoring operations. This layer transforms the abstract business requirements to applicable security mechanisms and specifications. In this layer, TOGAF-ADM presents EA principles and defines EA metrics and artifacts. SADM presents and prescribes the security-related EA principles, information security principles and information security services based on the security requirements and design principles. Also SADM presents security artifacts and security metrics. Eventually the integration of information security principles and a set of security related EA principles produces a list of quantifiable security metrics that permits information security architects to swiftly and easily assess the whole security of an enterprise based on the target artifacts, i.e., tracking attacks and their behaviors which are implemented on graphical diagrams presented using MAD modeling language. The security metric is one of the monitoring parameters of information security functions that provide evidences to independent auditors that intend security program is in place in an enterprise. Moreover, it is capable of measuring the level of the entire security of the enterprise with the help of the enterprise's architectural security-related artifacts.

The *Physical Architecture layer* (Builder's view) defines information security rules, practices, procedures and information security mechanisms. Specifications described by EA business rules constrained by information system standards including business practices and procedures are defined by the EA model. In this layer security mechanisms are associated with information security functions to achieve higher maturity levels of the enterprise such as policy documentation and metrics.

The *Component Architecture layer* (Tradesman's view) shows the implementation of components and the alignment they have with each other. It is the integrated configuration and deployment of the information security standards, products and tools in complement with the EA standards, products and tools instantly side by side.

Finally, the *Operational Architecture layer* helps to do monitoring and evaluation of the functioning enterprise as well as the information security service which are performing in the EA. It is the overall security service management of the organization.

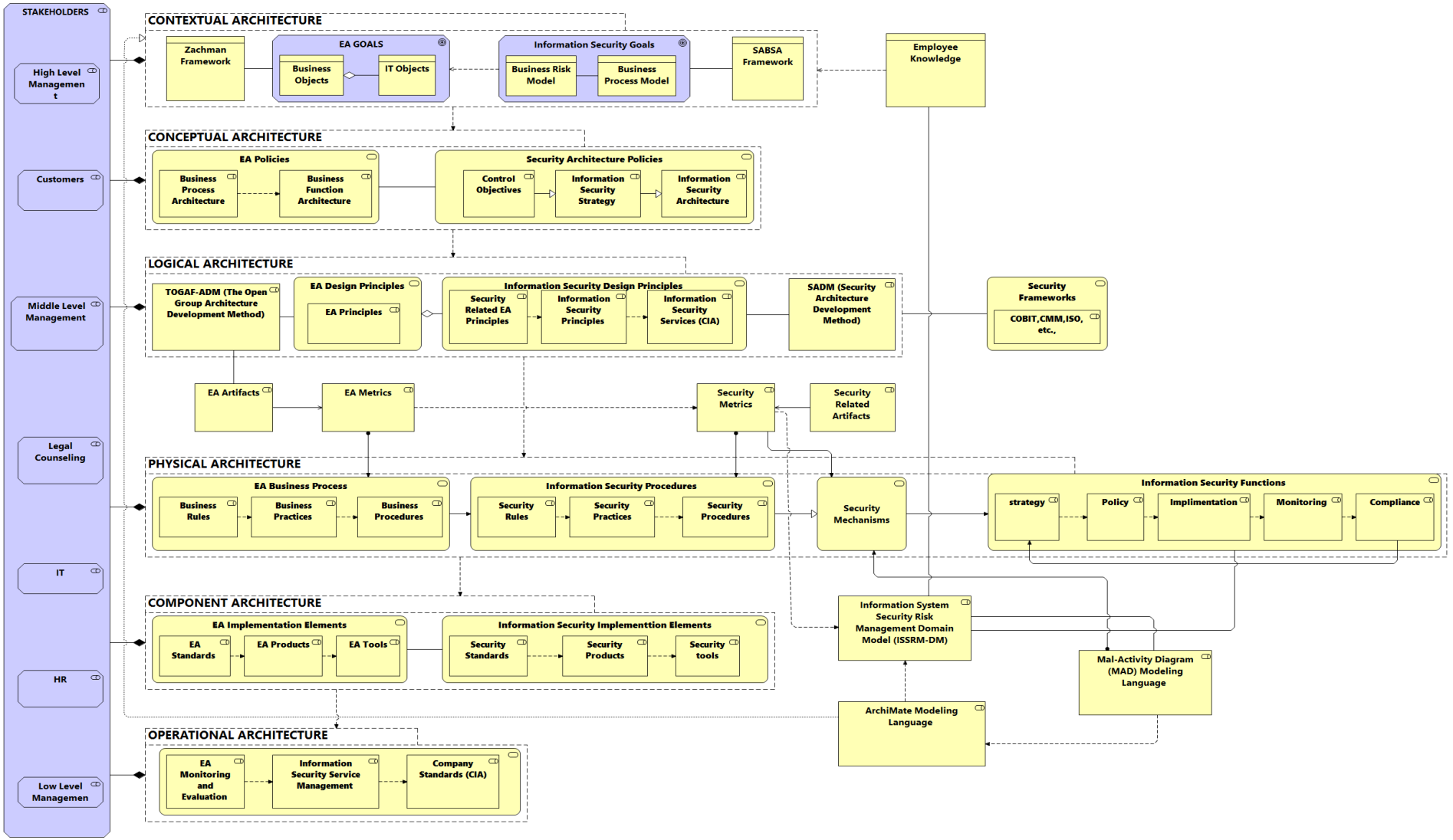


Figure 1: The Proposed Security-Integrated Enterprise Architecture Framework (SIEAF)

4. Discussion

The proposed framework is validated both internally and externally by reviewing the design process and expert interviews. Theoretical contribution of this paper has validated by research rigor. That is by the components of theoretical contribution which comprises what, how, why and the limitations of the theory by questioning who, where and when. Moreover, this framework was evaluated by relevant security and enterprise architecture experts. The evaluation is mainly on the usefulness and its degree of implementation from practical point of view. Overall the feedback was positive. However, two of the interviewees did not agree on using two modeling languages and a domain model that might result in a complex graph presentation of concepts, reducing the overview and vision. The remarks made by the experts have some implications for the suggested framework. To conclude, the approach is determined to be useful as well as usable, considering the highlighted limitations.

5. Conclusion and Future Work

The aim of the paper is to develop knowledge artifact in the form of a framework which focuses mainly on the integration of information security and EA at an initial stage of the EA development process. Information security and EA principles, which have a major effect on security, are the key ingredients of the integration.

The designed framework contributes to the currently inadequate theoretical background of knowledge, specifically to the field of EA and information security architecture and their integral relation, challenges and limitations for protection and management of business. The practicality of SIEAF should be considered relevant and valid, though the artifact still requires to be tested and validated in real environments.

In the future, studies will be carried out on application of mathematical model which quantifies the security level of each layer starting from its early EA development stage by considering EA and security design principles. The model will enable to design security sensitive enterprise assets by providing enterprise architects the ability to assess their organization automatically based on the security metrics of the enterprise.

References

[1] S. Jalaliniya and F. Fakhredin, “Enterprise Architecture and Security Architecture Development,” Unpublished Master Thesis, Lund University, 2011.

[2] B. M. Alshammari, “Enterprise Architecture Security Assessment Framework (EASAF),” Journal of Computer Sciences, 2017, 13(10), pp. 558.571.

[3] S. V. Bosch, “Designing Secure Enterprise Architectures,” Unpublished Master Thesis, Mathematics and Computer Science Enschede, the Netherlands, 2014.

[4] S. Bondar, J. C. Hsu, A. Pfouga, and J. Stjepandic, "Zachman Framework in the Agile Digital Transformation,” 2017.

[5] The Open Group, “Integrating Risk and Security within a TOGAF Enterprise Architecture,” Secur. Forum (a Forum Open Inst. Group), 2016, Available at https://publications. opengroup.org/g152.

[6] I. Tovstukha, “Management of Security Risks in the Enterprise Architecture using ArchiMate and Mal-activities,” Unpublished Master Thesis, University of Tartu, 2014.

[7] S. Larno, V. Seppänen, and J. Nurmi, “Method Framework for Developing Enterprise Architecture Security Principles,” Complex Systems Informatics and Modeling Quarterly, CSIMQ, No. 20, pp. 57–71, 2019.

[8] https://store.theartofservice.com/the-zachman-framework-toolkit.html, 2018.

[9] TOGAF Version 9.1, G116, The Open Group, 2011.

[10] Chowdhury M. J. M., Matulevičius R., Sindre G., and Karpati P., "Aligning Malactivity Diagrams and Security Risk Management for Security Requirements Definitions", 2012.

[11] V. Czaplewski CBAP, CMRP, SABSA Foundation SCF, 2017.

Modelling an Offline Electronic Cash Transfer Capability to the Existing Mobile Money Platform: Case of Social Cash Transfer Programs

Hussen Seid
HiLCoE, Software Engineering Programme, Ethiopia
mailtohussen@gmail.com

Mesfin Kifle
HiLCoE, Ethiopia
Department of Computer Science, Addis Ababa University, Ethiopia
kiflemestir95@gmail.com

Abstract

The Productive Safety Net Programme (PSNP) is Ethiopia's rural safety net for chronically and transitorily food insecure households by providing cash and/or food transfers. These interventions aim to strengthen resilience, improve nutrition, and help households become food enough and, eventually, food-secure [1]. The cash transfer program has evolved through various stages from manual cash disbursement by the government agencies to the electronic cash transfer which is facilitated by mobile money platforms.

The electronic cash transfer method played an important role in the disbursement of cash to the beneficiaries as it provides timely, convenient, and secure service. But, due to the cellular network coverage and consistency problem, it is not possible to address all beneficiaries in the program by the electronic cash transfer system within the mobile money ecosystem in Ethiopia.

This paper comes up with an operational process and a technological solution by designing a software system in which offline areas can be addressed by the existing mobile money platforms. This is realized by modelling a software system that can be integrated into existing mobile money platforms. The system has two main parts: the desktop application which can be used by users who have online access to the mobile money platform, and the POS (Point of Sale) terminal application which can be used by an offline agent. The offline and the online systems are integrated by sharing a cryptographic file/key and offline capability and the existing systems are integrated via a database.

Keywords: Social Cash Transfer; Mobile Money; Advanced Encryption Standard (AES*)*

1. Introduction

During the past two decades, much social cash transfer (SCT) programs have emerged in developing countries as a promising means for social protection. Since their introduction as pilots in Latin America in the early 1990s, the popularity and support of SCTs among national governments as well as the international development community have increased considerably. SCTs have, therefore, moved "from the margins of development policy towards the mainstream in several global regions" [2]. In general, the term 'social cash transfer' describes "regular non-contributory payments of money provided by government or non-governmental organizations to individuals and households" [3]. SCTs can be both conditional and unconditional and they are commonly targeted at households or persons fulfilling certain eligibility criteria such as income poverty or malnutrition. As the largest SCT programs are carried out in middle-income countries such as Brazil, South Africa, and Mexico, only a minority of beneficiaries originate from states with a low average income. However, recently, several SCT pilot programs and nationwide initiatives have also been implemented in Sub-Saharan Africa and other low-income regions [4] demonstrating that SCTs are now considered as adequate instruments of social protection for the least developed countries.

The main goal of a delivery mechanism (or payment system) within an SCT program can be described as "successfully distributing the correct amount of benefits to the right people at the right time and frequency while minimizing costs to both the program and the beneficiary" [5]. The available delivery mechanisms can be broadly divided into 'pull' and 'push' methods. Traditionally, most SCT

initiatives applied a manual 'pull' system, which requires beneficiaries to collect their regular cash transfer at a specified date and location. These pay points are oftentimes local government offices, mobile offices administered by local authorities or NGOs, or post offices [6]. It is due to these substantial weaknesses that many development agencies have recently abandoned manual 'pull' methods and started using innovative 'push' systems, which 'push down' delivery to the individual beneficiaries [7]. In this respect, particularly electronic delivery methods have gained currency, as they outperform manual methods in multiple aspects. Almost 50% of all SCT programs implemented in the past decade use this type of delivery mechanism [8].

Any electronic delivery method consists of two steps. First, governments or international donor organizations electronically transfer cash to a bank which deposits the benefits impersonalized accounts that can be conventional bank accounts or limited-purpose instruments such as e-money accounts. The second step consists of beneficiaries withdrawing cash from these accounts at a network of pay points. Several electronic delivery methods exist that differ in the network of pay points they use. Pay points can be bank branches or post offices, Automated Teller Machines (ATMs), contracted shopkeepers with POS devices, or mobile money agents [9].

The inability of mobile money platforms of Ethiopia to provide electronic cash transfer to the beneficiaries located at non-cellular areas is the motivational factor for this research to study the existing systems and design an operational process and a software system that can enable existing mobile money systems to support social cash transfer and disbursement with an offline context.

The designed software system has two parts: online and offline. The online one is the main system that is integrated with the existing mobile money system. It is accountable to access an account and transaction history of account holders from the existing system and creating an encrypted file as an offline payment order which will be later used at the offline spot. Moreover, the online system is accountable for online account reconciliation by importing offline disbursement file. The online account reconciliation involves the debit and credit transactions on the online mobile money platform via the database integration channel.

The offline software system is a POS terminal system that can be installed on personal computers, mobile phones, or specialized hardware which can be accessed by the payer cashiers. This system is integrated with the online part with an encrypted file. The software system is capable of importing offline payment order files and disbursing the cash and exporting the transaction history as a withdrawal synchronization request, which is reconciled with the online system.

2. Literature Review

Mobile money constitutes an alternative electronic delivery method for smart cards or magnetic stripe cards. Governments set up mobile accounts for beneficiaries at Mobile Network Operators (MNOs). Accounts are linked to phone numbers and the government provides the MNO with a list of recipients including their Personal Identification Number (PIN) and phone number. A text message informs beneficiaries that a payment was transferred, and funds are debited from the government's account. Beneficiaries can then cash out their transfers at mobile money agents by inserting their personal Subscriber Identification Module (SIM) card in any mobile phone and entering their PIN. Having received a confirming text message, the mobile money agent cashes out the desired amount. Mobile money agents are independent entrepreneurs charging beneficiaries a service fee for withdrawing money out of their mobile wallets. Beneficiaries cannot only cash out their benefits but also store and transfer money between different accounts (e.g., paying bills) or buy airtime and goods at participating vendors. Whereas some countries allow MNOs to store money on behalf of their customers, in other countries formal banks need to be involved [8]. Although mobile money has not yet been applied in large scale SCTs, successful pilots exist.

The 'Post Election Violence Recovery' program in Kenya used M-PESA to deliver cash to 37,000 individuals [10] and in Haiti, several NGOs relied on

mobile money tools for their emergency relief work in the aftermath of the earthquake in 2010 [8].

As governments can leverage existing mobile money agent networks, initial set-up costs for electronic delivery solutions using mobile money tend to be lower than for smart card-based systems. Governments do not need to provide contracted agents with POS devices or set up ATMs, yet they may need to provide beneficiaries access to a mobile phone if they do not possess one. In addition to financial transactions, mobile phones provide an opportunity for communication between the government and beneficiaries. They allow beneficiaries to provide feedback by sending text messages and facilitate governments to respond to complaints or to inform about changing conditions of transfers. Such two-way communication is likely to reduce leakage and improve accountability. Governments may also use text messages to influence beneficiaries' behavior. For example, beneficiaries could be encouraged to use transfers for school fees [8]. One disadvantage of mobile money systems is that they require constant network coverage, which is likely to pose a barrier in remote areas. Moreover, unlike smart cards, which only require beneficiaries to provide a fingerprint, mobile money solutions demand them to operate a mobile phone. Unless training and guidance are provided, this might constitute a barrier to access especially for illiterate individuals [11].

Like mobile money, mobile vouchers also work with MNOs. Yet, mobile vouchers require less hardware than a mobile money solution since only contracted shopkeepers need to be equipped with a mobile phone. Beneficiaries receive a PIN and a voucher number on two separate scratch cards and approach a contracted shopkeeper to redeem their voucher. They do not need to redeem the full amount at once and can usually only receive in-kind benefits [11]. To verify the transaction, shopkeepers enter the voucher ID and the redeemed value into their mobile phone and ask beneficiaries to enter their PIN. Upon successful verification, shopkeepers receive a text message and are then allowed to provide beneficiaries with goods and services. The redeemed value is credited to the shopkeeper's account. Alternatively, with electronic vouchers, shopkeepers verify transactions via the Internet. Therefore, they need to possess a smartphone or computer and access to a data connection to process this type of voucher [11].

Mobile vouchers are easy to set up, which makes them useful delivery methods for emergencies. For example, in response to the humanitarian crisis in Siri, the World Food Programme implemented an electronic voucher program in 2013, which allowed 800,000 refugees to buy food at local shops in Lebanon and Jordan [12]. Since vouchers usually only entitle beneficiaries to receive in-kind benefits, voucher redemption is not constrained by shopkeepers' cash flow. In contrast to paper vouchers, with mobile or electronic vouchers, shopkeepers can be reimbursed more quickly and governments receive more detailed data on beneficiaries' voucher redemption patterns. One main downside of mobile or electronic vouchers is that shopkeepers need network connectivity in order to verify transactions [11]. Moreover, vouchers are more suitable for short-term cash transfer projects since periodically new vouchers need to be issued.

Another alternative is a smart card. Smart cards are plastic cards with an embedded chip containing information on the beneficiary and transfer values. Recipients identify themselves at pay points (e.g., ATMs, post offices, agents with POS device) with a PIN and/or biometric identifiers and can then withdraw cash. Fingerprints and iris scans are mostly used for identification purposes. Which biometrics is most suitable depends on the beneficiary's occupation and age. Some jobs may cause damage to either hands or eyes and, in contrast to irises, fingerprints only stabilize around the age of fourteen. Moreover, cultural preferences should be considered. For instance, in Muslim populations, iris scans might be more culturally acceptable since no physical contact is required [13]. Smart cards have recently been introduced to large-scale SCT initiatives in several countries including South Africa's 'Child Care' and 'Old Age Pension' programmes as well as Mexico's 'Oportunidades' programme [14]. Furthermore, following successful pilots, Deutsche Gesellschaft für Internationale Zusammenarbeit

(GIZ) supported the Indian Government in the application of this technology in the national safety net program [15].

One of the three key advantages of a smart card is that no constant network connectivity is required, which makes smart cards particularly useful in remote areas. Since transaction information is stored both on the POS devices and the smart card chip, it is enough for agents to connect the device to the network at times to reconcile the accounts [8]. The second key advantage of smart cards is that information stored on the cards can be updated after the cards have been issued [9]. This allows governments to eventually streamline a range of services to citizens. For example, governments can easily add another 'wallet' to the smart card containing information on beneficiaries' eligibility to access other social protection programs, which deliver commodities (e.g., fertilizer) or services (e.g., subsidized education, health care) thereby convergence between different programmes targeted at similar population groups may be improved. Smart cards can also be used to store information on medical history or voting records [6]. Third, if smart cards rely on biometric identifiers, they are very easy to use even for illiterate individuals, as they do not require to memorize and enter a PIN or to own a mobile phone [10].

In addition to the above three alternatives, magnetic strip cards can be used for social cash transfer system. Magnetic stripe cards store data on magnetic stripes using iron-based magnetic particles on a band on the card. Beneficiaries identify themselves with a PIN or a signature, which makes magnetic stripe cards less secure and slightly more complicated to use than smart cards employing biometric identifiers. Beneficiaries can cash out their transfers at ATMs or contracted agents with POS devices. Magnetic stripe cards are used in some of the world's largest SCT programmes such as 'Bolsa Familia', which delivers cash to more than 12 million Brazilian households regularly [14]. Besides, they have occasionally been employed in emergencies, for example, in the aftermath of devastating floods in Pakistan in 2010 [8].

In contrast to smart cards, magnetic stripe cards need constant network connectivity, as transactions are performed in real-time. Another weakness of the magnetic stripe card is that no new information can be added once the card has been programmed. While these features make magnetic stripe cards an inferior solution compared to smart cards, the latter cause considerably higher set-up costs [11]. Smart cards are up to five times more expensive than magnetic stripe cards and chip-reading POS devices cost twice as much as those for magnetic stripe cards. Given these high initial investments for smart cards, governments should carefully consider whether the additional costs compared to magnetic stripe technology are justified. In particular, the offline capability of smart cards might become less relevant to increasing mobile network coverage [16].

This work is fundamentally different from other works. First, since it proposes a solution that will not require new full system development, rather it can be integrated into the existing mobile money platforms, no need for a mobile phone, SIM card, or cellular network to make a cash withdrawal. From the payer side, only an encrypted file and POS terminal device will be used without any cellular or Internet connection. Second, the reviewed systems work for developed countries with heavy bank and telecom infrastructures like ATMs, POS, and smart mobiles with high Internet and cellular network coverage. Beyond this, the proposed systems are mainly designed to facilitate merchant payments in highly banked and financially included society without concern of social cash transfer to support the livelihood of the people.

3. The Proposed Solution

The research revolves around developing a system that enables existing mobile money platforms to perform offline electronic payment or disbursement of social cash transfer. The system targets two main platforms. The first is adding offline transaction capability to the existing mobile money platform and the second is a POS terminal application that runs on a cashier's or agent's mobile phone to disburse the cash to the beneficiaries. The POS terminal application is served by an encrypted

payment disbursement file and encryption key generated from the central mobile money platform.

A. Process Design

The process starts with sending a request for offline activation of a certain location with the custodian of a certain cashier near to the offline group by a branch or donor to the central payment unit or disbursement unit. The request will be delivered or asked via modalities like e-mail, formal letter request, phone call, or SMS upon the agreement reached between stakeholders in the offline payment process. The stakeholders are branches, donors, government bodies, etc.

Upon the request of offline activation, the central disbursement unit, the one responsible to make a bulk electronic cash disbursement to specific beneficiaries at specific woreda or kebele, starts to group offline located beneficiaries in one offline group and generates an encrypted payment order file as an offline group and share the encrypted file to the requesting branch or cashier for offline cash redemption. In the meantime, the central disbursement unit also shares the decryption key for specific offline file with the payer cashier. The file and the key will be shared on any communication channel upon the consent of stakeholders in the process. It can be an e-mail, SMS, or delivery persons who can reach the location.

Then the offline payment process will follow. The cashier will import the offline payment order file to the POS system by providing the decryption key. The successful import will result in saving all necessary account information of a beneficiary and enable him/her to make cash withdrawal from the nearby cashier. Then the cashier will disburse cash in return for the deduction of a certain electronic amount from the beneficiary's account number. The beneficiary must visit the cashier nearby and provide the account number, PIN and amount to be withdrawn. A successful traverse on the system by entering the credentials, the cashier will give the reverse cash to beneficiaries as long as the credentials are correct and the amount to be withdrawn is less than or equal to the account balance on the system.

After finishing the offline cash disbursement or withdrawal from the offline cashier's POS terminal to the beneficiary, the cashier must export the withdrawal history of user accounts at his/her post and send the encrypted file to the central disbursement unit for online account reconciliation, getting all his/her money paid at the offline post to the real online account of the payer cashier and debiting or deducting the amount from all beneficiary accounts with the same amount withdrawn by the beneficiary at the offline post.

B. High-Level Design

The main mobile money system will have an additional module to support the offline electronic money transfer and development of a new POS terminal application that enables cashiers to make actual cash withdrawal with the beneficiaries based on the encrypted file generated by the main mobile money system/software.

The application needs to first cache data dictionaries and decrypt the crypto file before making the offline cash withdrawal. The POS terminal application, therefore, has its own data storage. The local database keeps track of all transactions that a beneficiary has withdrawn with the respective amount of transaction. The POS terminal application not only serves as a cashier station but also offers an interface to the beneficiaries to verify and authenticate their account and transaction.

As shown in Figure 1 [18], the centralised database also serves as a store for all configurations and data dictionaries. The main database is interfaced with a web-based or desktop application that enables system administrators and other console users to administer users, to make a transaction, to generate encrypted offline transaction files, and to reconcile offline transactions.

C. Algorithm Design

Algorithms are designed for critical operations of the system. The critical tasks are *account validation*, *user authentication*, *account import*, *account* export, *withdrawal export*, and *account reconciliation*.

Account Validation: The account validation engine deals with beneficiary authentication at withdrawal stage. The system will check the account number first and will proceed to PIN code checking if the provided account number is valid. If the

account number and PIN don't match, the system will not allow the beneficiary to withdraw the cash from the POS terminal.

User Authentication: The User Authentication engine deals with the payment specialist authentication or login process on the desktop application. The user must provide username and password to the system. Then the system checks the password and username combination against the credentials saved in the database. If one or two of the password or usernames failed to match with the settled values, the system will not guarantee access to the user interface.

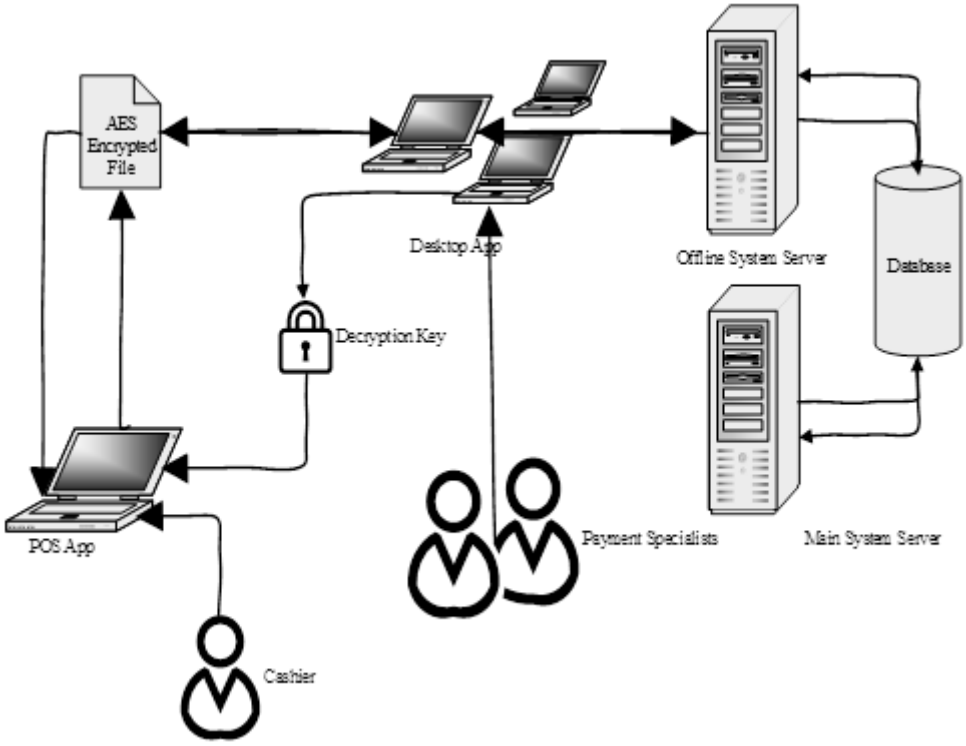


Figure 1: High-Level Design of the Proposed System

Account Import: The account import engine is about importing account information on the POS terminal application for the offline withdrawal process. The account information is imported from an encrypted file that is shared from the online system. The cashier must select the file and the decryption key. If the decryption key matches the key stored in the encrypted file, the account import process will continue, and the information will be stored in a local database on the POS application

Account Export: The account export engine is about exporting account information from the desktop application and encrypting the exported file. For user accounts to be exported for offline withdrawal, a cashier number must be provided to the system.

Withdrawal Export: The withdrawal export engine performs exporting user account information from the POS terminal application, writing the information to the file and encrypting it. The file will be used later for account reconciliation at the online system.

Account Reconciliation: The reconciliation engine is accountable for importing user accounts from the encrypted file, decrypting it, and updating the account information on the database system. This process is handled by the payment specialist on the desktop application with an online connection to the mobile money platform.

4. Prototype

To realize the feasibility and validity of the proposed solution, classes and methods are converted to source code. The prototype has two main parts.

A. Desktop Application

The desktop application is implemented using Java in NetBeans IDE. It follows object-oriented design patterns to structure the low-level functions. The software can run on an Operating System independent compiler and on a Java Virtual Machine.

As shown in Figure 2, the system offers capabilities that enable

- User Authentication,
- Offline group creation,
- Encrypted Payment order file creation and export,
- Withdrawal sync file selection and import,
- Financial transaction reconciliation, and
- Reporting.

Figure 2: Home Page of the Desktop System

B. POS Terminal Application

The POS terminal application is also developed using Java on NetBeans IDE. As shown in Figure 3, it provides an interface that enables cashiers to disburse electronic cash to the beneficiaries. The application offers capabilities that enable

- File selection and import,
- File decryption,
- File encryption and export,
- Beneficiary authentication,
- Electronic financial transaction, and Reporting.

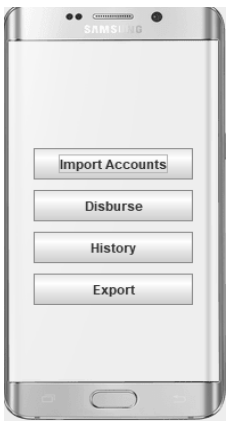


Figure 3: Home Page of the POS System

5. Conclusion

The electronic cash transfer on social cash transfer programs through mobile money platforms plays an important role in the disbursement of cash to the beneficiaries as it provides timely, convenient, and secure service. Beyond that, it calls for financial inclusion for the unbanked society.

The current mobile money technology does not serve social cash transfer beneficiaries living in offline locations due to cellular network coverage and consistency problem and a technology gap within the service providers in Ethiopia.

To alleviate the above gap, first, this research comes up with an operational process in which the non-mobile network areas can be addressed with the existing branch and agent networks of the mobile money service providers. Second, this research proposes a technolog by developing a software system model (a desktop system which serves users with direct access to the mobile money platform and the offline POS application which is used by agent and branch networks without direct online access to the existing system) that can be integrated with the existing mobile money platform. The offline and the online systems are integrated by sharing a cryptographic file.

The model is a new concept that also has several implementation challenges like the provision of training for stakeholders and integration with the existing system. The paper has identified operational procedure that can greatly enhance the social cash transfer program specifically to the non-cellular network areas and at the same time modeling and designing a software system which can be used for this purpose.

The algorithms and models used in the development of the system are not limited to the use case of social cash transfer programs. The same model can also apply in other tasks such as door to door revenue collection, shipping, and delivery.

References

[1] Ministry of Agriculture, "Productive Safety Net Programme Phase IV Programme Implementation Manual", Addis Ababa, Version 1.0, 2014.

[2] References AFI (Alliance for Financial Inclusion), "Mobile Financial Services: Basic Terminology", AFI Guideline Note, 2012.

[3] Samson, M., "Social Cash Transfers and Pro-Poor Growth", In OECD (Organisation for Economic Co-operation and Development), 2009.

[4] Bankable Frontier Associates, "Promoting financial inclusion through social transfer schemes", Report commissioned by UK's Department for International Development (DFID), Somerville, 2008.

[5] Grosh, M., del Ninno, C., Tesliuc, and E. Ouerghi, A., "For Protection and Promotion. The Design and Implementation of Effective Safety Nets", The World Bank, Washington D.C., 2007

[6] Devereux, S. and Vincent, K., "Using technology to deliver social protection: exploring opportunities and risks", Development in Practice, 2010.

[7] Vincent, K. and Cull, T., "Cell phones, electronic delivery systems and social cash transfers: Recent evidence and experiences from Africa", International Social Security Review, 2011.

[8] Smith, G., MacAuslan, I., Butters, S. and Trommé, M., "New technologies in cash transfer programming and humanitarian assistance", Report commissioned by the Cash Learning Partnership (CaLP), Oxford, 2011.

[9] Emmett, B., "Electronic payment for cash transfer programmes: Cutting costs and corruption or an idea ahead of its time?", Pension Watch Briefing 8, HelpAge International. London, 2012.

[10] Barca, V., Hurrell, A., MacAuslan, I., Visram, A., and Willis, J., "Paying Attention to Detail: How to Transfer Cash in Cash Transfers", Oxford Policy Management Working Paper 2010/04, Oxford, 2010.

[11] Sossouvi, K., "E-transfers in emergencies: Implementation support guidelines", Research commissioned by the Cash Learning Partnership (CaLP), Oxford, 2013.

[12] WFP (World Food Programme), "Syrian refugees get e-card to buy food in Lebanon", WFP news release, 11 October 2013.

[13] Gelb, A. and Decker, C., "Cash at Your Fingertips: Biometric Technology for Transfers in Developing Countries", Review of Policy Research, 29(1): 91-117, 2012.

[4] Bold, C., Porteous, D., and Rotman, S., "Social Cash Transfers and Financial Inclusion: Evidence from Four Countries", CGAP Focus Note 77, Washington, D.C., 2012.

[15] GIZ (Deutsche Gesellschaft für Internationale Zusammenarbeit), "A Smart Card Solution for India's Public Distribution System (PDS)", Eschborn, 2013

[16] Pickens, M., Porteous, D., and Rotman, S., "Banking the Poor via G2P Payments", CGAP Focus Note 58, Washington, D.C., 2019.

GTCM Enabled User Involvement in eXtreme Programming Approach

Mohammed Nuru
HiLCoE, Software Engineering Programme, Ethiopia
mohammed.nuru@aiesec.net

Mesfin Kifle
HiLCoE, Ethiopia
Department of Computer Science, Addis Ababa University, Ethiopia
kiflemestir95@gmail.com

Abstract

User involvement has been shown to be beneficial in eXtreme Programming projects. However, it has been a challenging issue in terms of engaging the right target of as onsite-customers. Despite researchers' efforts to improve the quality of user involvement in eXtreme Programming projects, user involvement gaps are still shown to be demanding a solution. We conducted literature and assessment-based study of the current practices of GTCM (Goal Target Channel Message) Concept and user involvement in eXtreme Programming and demonstrated the proposition of new approach and evaluations. The result of the research is an approach for the GTCM enabled user involvement in eXtreme Programming for the selection of appropriate users to participate in software development projects. The study used a design science research approach triangulated with qualitative and quantitative assessments and evaluations through action research and expert opinion to address the research objectives.

This paper proposes a new approach to improve the quality of user involvement in eXtreme Programming projects by using the GTCM Concept for the selection of users to participate in projects. The approach is evaluated in two iterations of project-based action research study and expert evaluation. The results show that the approach improved the quality of user involvement as well as user satisfaction. It is envisaged that the results of the research can be applied in any eXtreme Programming project to achieve the quality of user involvement.

Keywords: User Involvment; Agile Development; GTCM; eXtreme Programming; Software Engineering

1. Introduction

Today, software is eating the world [1]. There are no technical or organizational processes and services that are not related to software systems. The communication and lifestyle of people depend to a large extent on software systems. The security and critical infrastructure of nations depend on software and software engineering practices play a large role in the industry. The largest and most successful companies around the world are now software companies.

Software engineering has always suffered from fashions. Each decade in the past years had its own technology fashions including the disciplines and practices of software engineering. But as always, the truth is never in just one approach. The old saying “there is no silver bullet” [2] is still valid again so in fact, the agile approach is also not the silver bullet.

However, 78% of organizations practice Agile approaches [3] and the number is growing every year. One of the most widely used approaches of agile philosophy in the software engineering industry today is the eXtreme Programming approach. Sixty four percent of organizations use it with the adoption of other approaches [3].

In practice, one of the signs for the best quality of a software product is users' satisfaction [4] and improvement in user satisfaction has direct correlation with the quality of user involvement in the project [5]. Half of the software projects are successful due to their 52% focus on user satisfaction. Basically, satisfaction of users can also be used as one of the measurement aspects for software projects' success [3]. It is proved that user involvement positively affects user satisfaction [6, 7] and many researchers have investigated a way to increase user involvement in agile and also in eXtreme Programming approach.

In organizations working with people as their market engine, the participation of people is a critical strategy for organizational success. Such organizations realize a 50% higher productivity [8]. GTCM Concept has been implemented in various projects as a key ingredient for such success and can be used in activities involving people and can be implemented for various software engineering projects.

The motivation for this research is the existing challenges in achieving quality software in the eyes of user involvement in eXtreme Programming. There is a huge gap of user involvement in eXtreme Programming. Despite efforts of researches in implementing different approaches for efficient user involvement in eXtreme Programming projects, the participation of inappropriate users has been a challenge [9], negatively affecting a project's overall success. Because of these challenges, there is a need for enhancing user involvement practices in the eXtreme Programming approach.

This study has two major research questions to answer: What are the challenges of user involvement in the eXtreme Programming approach? How can the GTCM Concept address user involvement in an eXtreme Programming approach?

This study used Peffer`s design science research approach [16] triangulated with Action research and expert opinion techniques gearing qualitative and quantitative methodologies for data collection and analysis. The research was conducted in three phases namely: current practice assessment, application of the approach, and expert evaluation.

2. Related Work

The authors in [10] presented a theoretical study focusing on the comparison of RUP and eXtreme Programming features. It discussed and compared the activities of both methodologies during phases of a primary investigation, analysis, design, implementation, and transition phases. It is claimed that the proposed model is more efficient than eXtreme Programming and RUP exploits human experience during software development without any practical case study. Such claims need to be evaluated by different model evaluation techniques and also through feedback from the industry.

A novel process model SPRUP was proposed by integrating eXtreme Programming, Scrum and RUP and was validated through a controlled case study in [11]. It was claimed that the integrated model will be efficient, effective and fulfills the needs of business requirements. But the authors did not conduct a comparison of the proposed model with other process models or at least evaluated through case studies because the usability and effectiveness cannot be measured without comparison and empirical evaluation in which, and according to [12], the evaluation is supposed to describe how closely it scores on a test corresponds (correlates) with behavior as measured in other contexts.

Technical reports [13] show that GTCM adopted the 4 Disciplines of Execution model (4DX) to plan the digital marketing section and aimed to increase the project goals and they conducted an evaluation of the section growth which indicated the achievements the organization made on its programs against the year before and they scored 482.72% and 150% growth in two of their goals.

The authors in [14] tried to investigate the challenges that emerge during user involvement in eXtreme Programming, analyze and propose a solution. The work addressed the benefits of involving users and the problems that may drag the project into failure. The authors attempted to investigate the challenge that the projects have experienced from users who are not active and available, part-time users, lack of communications, change in requirements, and indicated the possible solution. The authors suggested Product Management Team to act as a user during requirement elicitations. The finding did not indicate any scientific methodology that will support the technical team to identify the most appropriate users to participate in the technical team that can replace part-time users by fulltime. According to [15], a technical team is not recommended to take part and influence the requirement elicitation process during system development. It will be a cause in terms of creating favoritism towards the needs of the technical team.

3. Current Practice Assessment Findings

Results from this survey suggest that GTCM Concept practices in the offices are inclined in three

major areas: project campaigns and promotions, selection and recruitment, and ICT.

Evaluation on participants' profiling and their exposure to projects [Figure 1] shows that the majority of the participants (76.9%) are exposed to marketing portfolios and projects, 65.4% to Talent management portfolios, 61.5% to ICT, 26.9% to exchange project participants, 19.2% to external relations participants, and 15.4% to finance and legalities participants. This exposure can be used as an indication of the experience levels of participants to ensure their appropriateness for the survey.

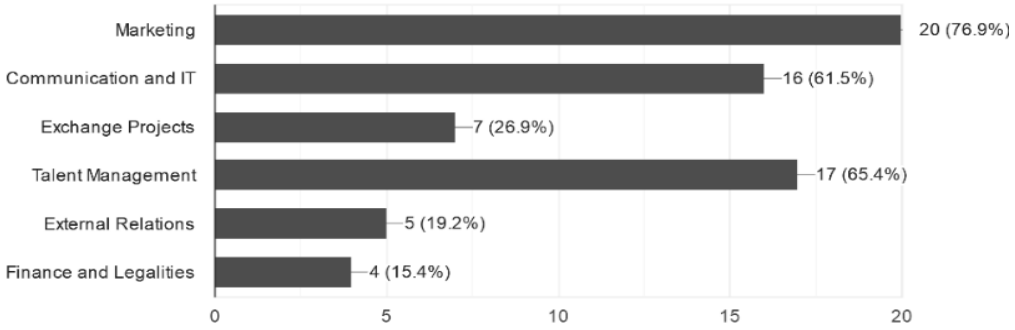


Figure 1: Project Activities GTCM is used for

The participants experienced the GTCM Concept in various portfolios and departments [Figure 2]. The majority (57.7%) of the participants used it for project campaigns, 53.8% for target user selection, ICT (50%), and recruitment related activities (34.6%).

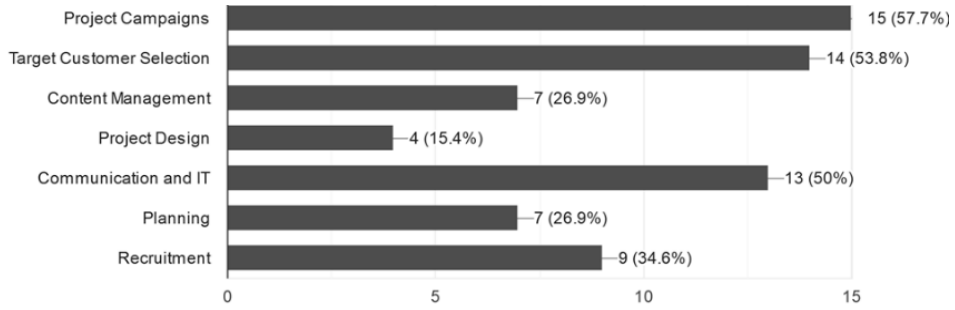


Figure 2: Participants' Exposure

Evaluation of the usefulness of the GTCM Concept, its exposure, and applicability in the IT industry is investigated and based on the experiences of the participants, 76.9% of them believe that the GTCM Concept is very useful and 23.1% think it is fairly useful. The majority (76%) of participants were engaged and exposed to the development of technological solutions using the GTCM Concept.

It has emerged that the GTCM Concept is very useful and can be used as a good practice. Furthermore, the concept can be used in IT projects in practices related to planning, recruitment, branding, product development, communication, and campaigning.

On the other hand, results from the interviews show the current practices of eXtreme Programming experienced a significant challenge during user involvement. The participants mentioned challenges related to lack of technical knowledge, documented, organized and clearly defined business processes and regulations. All of the participants experienced the involvement of inappropriate users. Unavailability of users, even if they are available they are not actively participating and they also lack the mandates and ability to decide on some of the scenarios and in general their lack of interest in the domain, all extended from the inappropriate selection of participants and also sampling techniques because they also observed employees left out from the selection process.

Other challenges were also mentioned by the participants such as communication gaps observed during team meetings between developers and user participants. Lack of knowledge and interest in the domain is the major cause of communication gaps [17].

Similar perspectives are addressed in [15] with major indications to the limitation factors related to lack of capable and knowledgeable users participating in the development which is mainly caused by the inappropriateness of users for the role. Also, it was mentioned that a communication gap between users and developers is a major challenge during user involvement.

The projects did not use any selection criteria. Instead they select based on their monetary interests, closeness to the departments and their social acceptance. Because of these, they suffered many challenges during the selection processes. The participants mentioned the challenges that are related to weak selection process are due to inconsistent candidate evaluation, struggle when making selection decisions, lack of criteria and role definitions, administrative influence on both selectors and participants by using incentives and administrative powers. According to participants, implementation of new selection practices will have a positive effect on the process and resolve the challenges in terms of selecting the appropriate participants, support for participants' evaluation with common guidelines and methodologies.

4. The Proposed Initial Model

The proposed initial model can be characterized as an iterative approach. The eXtreme Programming process components are not changed during the development of the model, rather a new practice is injected into the planning stage as shown in Figure 3.

The planning phase is the first phase of the life cycle, where the adoption of the GTCM concept is extensively permuted. The project manager begins the cycle by building the plan, time, and costs of carrying out the iterations, and individual developers sign up for iterations. Then s/he develops the GTCM definition for the specific iteration project, the so-called selection procedure. The selection process is the key phase in the project because it is the stage where the backgrounds of the candidate participants are mapped against the goal of the iteration. The GTCM Concept is used as a tool to guide the selection procedure. The concept has four major components: *Goal*, *Target*, *Channel*, and *Message*. In this case, the concept is used for selection purpose and the main focus of the selection process is to finally deliver a list of individuals as targets. Therefore, the *Target* component of the concept is permuted to the final stage of the GTCM concept processes.

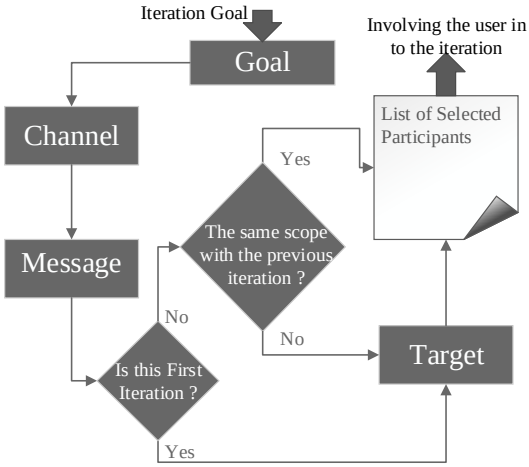


Figure 3: GTCM Process Model

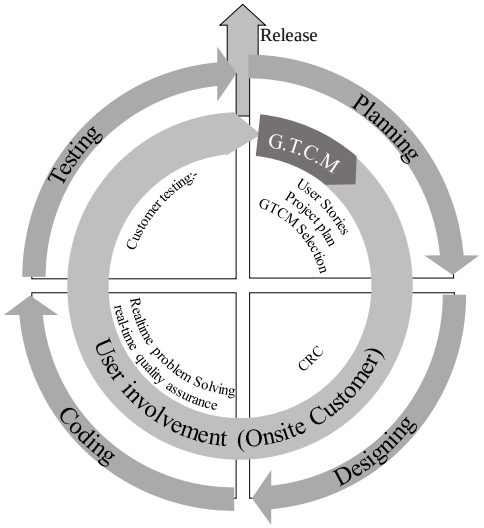


Figure 4: The Proposed Initial Model

The procedure will only guide the project team on the selection of participants for involvement in the project. The adopted selection procedure of the GTCM Concept is shown in Figure 4.

Step 1 - Goal Setting: Selection goals are project goals defined during the project charter with the project manager and stakeholders are directly used for this step. The goal is set for the purpose of letting other components to define their scopes and help the project manager to come up with a variety of criteria as per the need in order to achieve the selection goal.

Step 2 - Setting Channel: Channel is defined as a method and strategy of delivering a project, which will answer the *How* of the project. Software projects have their own ways of implementing projects in terms of project management methodology that the project manager chooses. A technology that is used to develop the system or any organizational standards that is going to affect the involvement of the users are mentioned and listed out at this stage. This level will help the selectors to identify participants based on their technical knowledge and backgrounds.

Step 3 - Setting Message: Message is the content of a project which defines the *What* of the project. Each iteration has its own functionalities and features expected to have at the end. This step details the description of the required high level functionalities of the system. It will help the selector to identify the work domain of the participants.

Step 4 - Target Selection: At this step we already know the *Goal*, the *What* and *How* of the project. Based on the settings, the selector will define the criteria for selecting candidate participants. Criteria are defined separately for each setting and are evaluated based on their predefined weightings by the selector.

The involvement level and range of users will depend on the criteria set in the components. Meaning, the required metaphor, and the definition of users are defined at this stage. *As an example, if the goal of the iteration is to develop a module on financial management activities, the end user will more likely be a finance expert with required experience in such a way that the channel and message are defined in a similar manner to select the best financial expert in the domain*.

Step 5 - Selection: At this step, users' profile is collected and sorted with the required information based on the criteria. Each candidate is evaluated and measured on a percentage scale. The top-scoring candidates are considered as the ideal participants that are going to be involved through the Onsite Customer practice in the selected iteration project.

Selected participants meet the development team to create ‘user stories’ or requirements. The development team converts user stories to iterations that cover the functionalities or features required for a single iteration and their combinations at the end provides the customer with the final fully functional product.

5. Application of the Approach

The results of the current practice assessment are used as input for this action research. Along with that, initial contact with the key participants was facilitated. One of the major activities conducted was the preparation of the project proposal of the action research. Secondly, a couple of meetings in the department were steered and raised the purpose and awareness of the research and support for the project. Based on the meetings and previous assessment output, facilitated discussion on the problem situation and intended actions have conversed and finally the current state is analyzed.

Iteration and further problem situations and outcomes are identified. Assessment of the team identified the potential team members to support and implement the proposed model procedures. Based on the number of team members, task identification and mapping were conducted with respect to the team members' roles in the department.

A project is identified based on the department's annual plan. From the selected project, the project team and stakeholders made a list of all the project iterations based on their priorities made by the stakeholders. For this study, two iterations were selected to be included in the action research.

Based on the initial proposed model, implementation of the selected iteration projects is conducted with the application of the procedures and practices of eXtreme Programming. The teams followed the proposed process model at the planning phase. Based on the three components of the GTCM Concept, the team have come up with a list of criteria for the selection and conducted the selection.

Users' profile are collected and sorted with the required criteria information. Each candidate's profile are evaluated by the assigned team members and sent a notification to the project manager. Accordingly, the project manager announced the selected participants to the relevant offices and stakeholders then proceeded to the following phases of the project.

After selection is finalized, the team constructed from both technical and users performed the planned activities and tasks. Since participants are part of the planning, they consider themselves not as users but also contributors to the success of the system. The participants contributed in a way defined by the eXtreme Programming practices and scopes.

The project team conducted an assessment each week and phases of the iteration. Meanwhile, the project result is evaluated using an NPS (Net Promoter Score) score (Table 1) to validate if the challenges and problems are resolved by cross-checking user satisfaction.

Table 1: User Satsfaction Evaluation (NPS)

Metrics	*1st iter*	*2nd iter*	*Diff*
User Story Correctness	69%	85%	▲16%
User Story Coverage	69%	85%	▲16%
User Story Completeness	85%	85%	0%
User Story Acceptance	85%	92%	▲7%
User Story flow	77%	85%	▲8%
Communication Level	77%	100%	▲23%
User interests	100%	100%	0%
Efficiency	75%	75%	0%
Influence reduction	75%	100%	▲25%

The result collected from stakeholders of the project indicates that assimilating the GTCM Concept into the eXtreme Programming approach as a user selection methodology has a significant benefit in improving the quality of user involvement as well as the project result. In addition, based on the comments, feedbacks and observations in both iteration implementations of the approach, improvements are made on the model.

The action research provided practical experience-based information on how the new approach should be used in projects. The study also offered an insight into user involvement matrices. The contribution of this part of the study is to guide the development of the model by understanding real practices through implementations and realizing comments from practitioners, and in the end, a more realistic and applicable model can be developed. The final improved model is shown in Figure 5 with consideration of the experts' evaluation outputs.

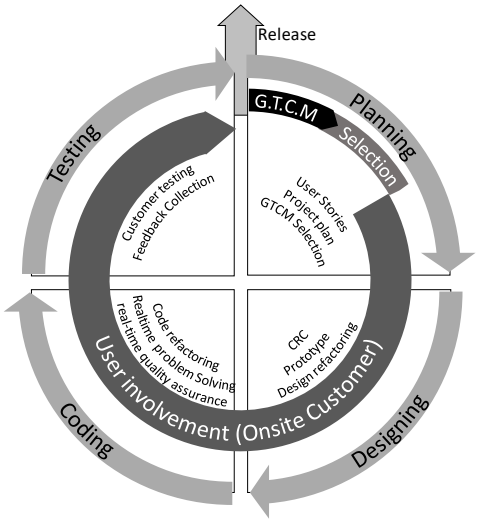


Figure 5: Final Model

6. Expert Evaluation

The main purpose of involving specialists in this study is to evaluate the improved model through the eyes of experts to get a better understanding of the model through critics, feedback, and recommendations to ensure the acceptability and applicability of the model in the industry.

Therefore, profiling of experts is conducted in three major qualification scopes as suggested in [19]. In accordance with the aims, resources, and questions of the evaluation, we tried to answer what is required of knowledge priority, knowledge depth, and the area of expertise. The second scope defines which approach is suitable for the evaluation (in accordance with different categories of experts and expert availability) in such a way to address which perspective, which affiliation, and how to convince. Finally defining who is/are suitable for the evaluation (in accordance with available published information about the experts and the willingness of experts to cooperate in other personal assessment methods) addressing the expert number, experts' knowledge, and experts' attitudes.

Five experts were selected for expert evaluation and their demographic diversity, experience, and backgrounds in agile philosophy and the GTCM approach are considered during the selection.

Primary data is collected through interviews. Five evaluation dimensions are used as evaluation technique [20]: *Goal*, *Environment*, *Structure*, *Activity*, and Evolution. *Goal* defines the efficiency, generality, and validity of the artifact scaling the level of correctness when producing the desired

effect. *Environment* consolidates the surrounding people, organization, technology along with their consistency with the people, organization and the technology. The *Structure* of artifacts is assessed by completeness, simplicity, clarity, style, and homomorphism. *Activity* is characterized by completeness, consistency, accuracy, performance, and efficiency. Finally, *Evolution* is characterized by robustness and learning capability of detail and consistency. After evaluation is conducted, key findings from the interviews are individually framed and thematically analyzed.

6.1 Findings

Goal: The experts agreed that the artifact can produce its desired effect on user involvement and it is goal driven and gives a value for users which will create a better extraction of user stories. However, the success of the approach depends on the accuracy of the goal set for the project and collecting feedback from users is a key activity that must be mentioned in the model besides user testing.

Environment: The approach can be considered as consistent with conditions that project managers should be aware that there must be a proper guideline, criteria and proper channels are used when selecting users. The approach creates an adoption between information and technology and connects people with the organization. This will improve the consistency that will enable us to change technological methods without loss of information.

Stricture: Its reflections through the models are complete, simple, and clear. However, suggestions are made to the models that definitions should be placed in the GTCM process to clarify the processes to new readers who are novice to the approach.

Activity: The proposed approach is applicable for dynamic project environments. It is generic and flexible in such a way that it is adapted to any software project environment with intentions to have quality software by involving users for improving their contribution to the development of the software.

Evolution: The ability of such models to be maintained at a certain quality or level is an important aspect of evaluation. Sustainability of the artifact maintaining its simplicity and completeness is evaluated during the interview. The experts approved its sustainability with a reason that the approach will give the project managers the ability to hold quality users in dynamic ways and further improvements on the approach can give a better result.

6.2 Suggestions and Recommendations

Advises and suggestions made by the experts on the overall stand of the artifact is discussed during the evaluation and summarized as shown in Figures 5 and 6.

Improvements can be made during the implementation of the model by improving the way to reach potential users, using surveys. It is challenging when the user domain is large. In that case it is recommended to use other techniques. The sensitivity of Selection should be considered when publishing the model to public use because it will directly affect the quality of the software.

According to the Evaluation results, the model is considered as a significant contribution to the improvement of user involvement in eXtreme Programming and Agile methodology in general. The findings indicate that it is possible to use the approach for software development projects in the industry and improve user involvement as well as software quality. The results attained from the evaluation reviewed the suggestions, recommendations, and explored the implications of the findings in relation to the evaluation dimensions.

7. Conclusion and Future Work

Both action research and expert evaluation results show that the artifact can provide a good result if implemented in projects. The findings may help to promote success in software projects, improving quality, delivering to scope, promoting timeliness and reducing the cost of eXtreme Programming software development projects.

Further investigations and research could be performed on different process model setups and business cases. Implementation of the artifact with other business processes to further address user involvement challenges and using methodologies out of the software domain to address the behavioral aspects of users to improve user satisfaction could provide a new dimension of addressing user involvement. Other scientific factors that will affect user involvement and overall project success will be considered during further studies. Factors may

include the number of user participants in the project, knowledge and experience of the project team, application of project management techniques, and implementation environment.

The applicability of the approach with the permutation of evaluation techniques to improve the appropriateness and perfection of the selection criteria used in the GTCM process is also a future work.

References

[1] J. C. Hollar, "Why Software is Eating, and Changing, the World," Computer History Museum, 2012. https://computerhistory.org /blog/why-software-is-eating-and-changing-the-world/, Last Accessed 2019.

[2] M. Chesbrough, "No Silver Bullet - why agile is not the answer," A Medium Corporation, 2018. https://martinchesbrough.net/no-silver-bullet-why-agile-is-not-the-answer-21e4b1bcba22, Accessed 2019.

[3] C. VersionOne, "13th Annual State of Agile Report," European SAFe Summit, The Hague, Netherlands, 2019.

[4] S. Hussain, S. Farid and I. Mumtaz, "Is Customer Satisfaction Enough for Software Quality?", International Journal of Computer Science and Software Engineering (IJCSSE), Vol. 8, No. 2, pp. 40-47, 2019.

[5] S. Kujala, M. Kauppinen, L. Lehtola and T. Kojo, "The Role of User Involvement in Requirements Quality and Project Success," in 13th IEEE International Conference on Requirements Engineering (RE'05), Paris, 2005.

[6] S. Kujala, "User involvement: A review of the benefits and challenges," Behaviour and Information Technology, Vol. 22, pp. 1-16, 2003.

[7] Z. Taha, H. Alli and S. Abdul-Rashid, "Users' Requirements and Preferences in the Success of A New Product: A Case Study of Automotive Design," Applied Mechanics and Materials, 2012.

[8] I. Gallup, "Foster Loyal Customers," Gallup, Inc, 2019.

[9] B. N. a. S. S. S. Mohammadi, "Challenges of user Involvement in Extreme Programming projects," International Journal of Software Engineering and its Applications, 2009.

[10] N. H. a. D. K. K. Fertalj, "Permeation of RUP and XP on small and middle-size projects," in Proceedings of the 5th WSEAS international conference on Telecommunications and informatics, Croatia, 2006.

[11] S. U. N. Rizwan, M. Rizwan and M. R. J. Qureshi, "Empirical Estimation of Hybrid Model: A Controlled Case Study," International Journal of Information Technology and Computer Science, Vol. 4(8), 2012.

[12] A. Negi, "Industrial Psychology, Lucknow, India," 2010.

[13] V. Soni, "AIESEC Ahmedabad Marketing 2015 Excellence Awards Application," 2015. https://www.slideshare.net/vhdvivek/aiesec-ahmedabad-marketing-2015-excellence-awards-application.

[14] S. Mohammadi and S. Sohrabi, "An Analytical Survey of On-Site Customer Practice in Extreme Programming," in Computer Science and its Applications, 2008.

[15] H. S. a. R. Dale, "Requirements engineering: Making the connection between the software developer and custome," in Information and Software Technology, Vol. 42, pp. 419-428.

[16] K. Peffers, C. Gengler, T. Tuunanen and M. Rossi, "The design science research process: A model for producing and presenting information systems research," Proceedings of First International Conference on Design Science Research in Information Systems and Technology DESRIST, 2006.

[17] SkillsYouNeed.com, "Common Barriers to Effective Communication," SkillsYouNeed.com, https://www.skillsyouneed .com/ips/barriers-communication.html.

[18] S. Mohammadi, B. Nikkhahan and S. Sohrabi, "Challenges of user Involvement in Extreme Programming projects," International Journal of Software Engineering and its Applications, 2009.

[19] B. Ramadurai and N. Becattini, "Classification of technology forecasting methods: strengths, weaknesses and integrability analysis," The FORMAT Project, 2013.

[20] N. Prat, I. Comyn-Wattiau and I. Comyn-Wattiau, "Artifact Evaluation in Information Systems Design Science Research – A Holistic View," in Pacific Asia Conference on Information Systems, 2014.

Oilseed Leaf Disease Identification Using Image Processing

Rahel Bekele
HiLCoE, Software Engineering, Ethiopia
Information Network Security Agency
titarichel@gmail.com

Solomon Gizaw
HiLCoE, Ethiopia
Department of Computer Science, Addis Ababa University, Ethiopia
solohavi@gmail.com

Abstract

Oilseed is the most important food crop which is gaining increasing demand in the world market. Though the demand increases in the world market, there are challenges to its production. Diseases are among those challenges that resulted in crop yield loss. The diseases can be caused by bacteria, viruses, fungi or other factors. Diagnosis of diseases is traditionally done by eye inspection. This way of identifying diseases is less accurate, very time consuming, more laborious and can only be done in limited areas. Shortage of domain experts, especially in developing countries, is the major problem in early identification and management of these diseases. In addition, estimation of disease severity is prone to subjectivity.

In this paper, we propose a solution to identify oilseed crop diseases that requires less time, less effort and at the same time provides more accurate detection. The proposed solution uses image processing techniques for identification of the diseases and grading of disease severity. Grading of disease severity enhances the identification processes by providing the information which helps for taking appropriate measures or treatments depending on the disease severity.

A prototype is developed using Matlab. Leaf images of oilseed, i.e., soybean and sunflower, are collected. K-means clustering technique is implemented for segmenting the images. Then color and texture features are extracted. For identification of diseases, multi class Support Vector Machine (SVM) is used. After identification, severity grading of the disease is done using fuzzy logic approaches. The prototype is tested and diseases are identified with accuracy of 97.2% and the stage of the diseases is properly predicted. This result is more effective compared to other reviewed works that achieved accuracy of 95% and 93.5% for specific oilseed crop leaf diseases.

Keywords: Plant Leaf Disease; Image Processing; K-Means; Feature Extraction; Multi Class SVM; Fuzzy Logic

1. Introduction

Agriculture where more than 60% of the world's population depends up on remains the center of the world economy [1]. The dependency is even higher in developing countries, which is a matter of survival. It is the major source of food, income, and employment in addition to its major contribution to Gross Domestic Product (GDP) [2].

However, the potential for the production, productivity and market opportunities of oilseed is increasing [3, 4], although a number of factors stand against the quantity and quality of the yield. For instance, lack of adequate knowledge of farming, insect pests, plant diseases and limited number of experts are among the major challenges. Particularly, diseases are causing significant yield loss [5, 6] which result in social, ecological and economical loss to farmers [7].

Furthermore, inadequacy of domain experts remains a problem with the sector especially in developing countries. For instance, Biruk Ambachew [5], who conducted knowledge based oilseed disease diagnosis research in Ethiopia, claimed that the experts are not always available and may not be accessible to every farmer. Each agent has to serve on average more than a thousand farmers [5]. Moreover, farmers and even agricultural experts might fail to identify the diseases precisely [7, 8].

Identification of plant leaf diseases using image processing techniques will mainly reduce the confusion experts and farmers face during diagnosis

of diseases, especially when uncommon diseases for specific areas breakout. It needs less effort and less time. It also enhances remote sensing of the diseases when experts are not available in the field.

Leaf carries feature information that enables to recognize the plant [25]. It is stated in [25] that each leaf of a plant has its own features, including oilseed crop leaves.

Works such as [26] have incorporated severity grading for Maple and Hydrangea leaves. As mentioned earlier, specific solution for specific crop has its own impact for precise recognition of diseases.

Though several researchers proposed techniques for leaf disease identification and provided significant result, they all have certain limitations [27]. According to Dhaware and Wanjale [27], to overcome the drawback of different techniques fusion of different techniques is a good idea. So in this paper we proposed best hybrid of segmentation, extraction and classification techniques for identification of oilseed crop leaf diseases and grading of the disease severity. The proposed solution will be more precise, requires less time and effort for diagnosing of diseases. On the other hand, grading of disease severity has high impact on determining the appropriate measure for mitigating the diseases. This will reduce massive yield losses and plays specific role in sustainable productivity of oilseed crops.

We designed a prototype for identification of oilseed leaf diseases using image processing technique. Fuzzy logic based approach is implemented for enhancing disease identification by pinpointing the stage of the disease severity. Fuzzy logic is multi-valued logic. The logic has more truth values than the classical logic which has only ‘true’ or ‘false’ (0 or 1) values. Incorporation of these values, between true and false or 0 and 1, enables to handle uncertainties. Hence, it results in better accuracy [13, 14]. “The uncertainties within image processing tasks are not always due to randomness but often due to vagueness and ambiguity. Fuzzy technique enables to manage these problems effectively” [14]. Fuzzy image processing in general provides a better solution to deal with uncertainties or vagueness and it is appropriate technique for expert knowledge representation and processing [14, 23] like oilseed disease severity grading which is prone to subjectivity. Such problems are difficult to resolve properly by only limiting the domain of analysis within the normal crisp set. However, representing of those uncertain or ambiguous data by allowing some degree of membership in a given set, in this case disease severity stages, will increase precision.

2. Related Work

Jakjoud *et al*. [10] proposed a solution for detection of tomato leaves disease. K-Nearest Neighbors (KNN) and SVM are used for classification of diseases. Fuzzy logic based decision is made for the classification in both approaches. The fuzzy interference system in this work contains if-then rules for identifying the healthy leaf from the infected one. Combined approach of SVM and KNN with fuzzy logic gave a better result than the global SVM and KNN. Overall accuracy of 98.38% and 82.01% have been achieved by Fuzzy combined KNN and Fuzzy combined SVM, respectively Though accuracy of 98.38% is very good, the proposed solution is limited to identifying the diseased one from the healthy one so it supports only two classes, i.e., it does not support classifying to specific disease types and severity grading.

The authors in [12] also proposed fuzzy logic based cotton leaf disease detection system. Gray-Level Co-Occurrence Matrix (GLCM) is used for extracting features and fuzzy logic for classification of diseases. An accuracy of 88% is achieved. Although cotton is one of oilseed crops, the proposed solution is limited to identification of cotton leaf diseases, i.e., it does not consider severity grading of the disease.

Gui and Mbaye [9] proposed identification of soybean leaf spot diseases using deep convolutional neural networks. The soybean images are downloaded from the Internet. Those images are then segmented using unsupervised fuzzy clustering algorithm. The proposed classification model is defined using Deep Convolutional Neural Network (Deep CNN). With this model accuracy of 89.84% is

achieved. The researchers have also compared the proposed model with SVM and Visual Geometry Group (VGG 16). As a result, VGG 16 achieved a testing accuracy of 93.45% while SVM achieved 83.23%.

Liu *et al*. [28] proposed identification method of sunflower leaf disease based on Scale-Invariant Feature Transform (SIFT) point. In this work, sunflower leaf images are acquired by mobile phones. After preprocessing the collected images, SIFT algorithm is used to extract feature points of the images. Feature extraction of sunflower leaf scale consists of processes like extreme value extraction, accurate positioning of key points, key point direction determination and generating SIFT feature vectors. Identification of sunflower disease is held by matching of image feature vector. An average accuracy of 95% is achieved in recognition of four diseases.

A knowledge based system is proposed for diagnosis of oilseed crop diseases in [5]. The system has 'yes' or 'no' questions that represent the symptoms of different diseases. If all the symptoms are observed or all the answers given by the user are 'yes' then the system displays the type of the disease diagnosed and it suggests treatments for the identified disease if the user needs. It works based on the knowledge already fed to the system. It does not analyze the disease from the leaf image itself. In addition, the questions are prepared in human understandable format. As a result, some characteristics or features of the crop required for disease diagnosis remain unkempt.

3. The Proposed Solution

The proposed oilseed crop leaf disease identification system architecture has three major components as shown in Figure 1. These are *training*, *testing* and *grading* the stages of the disease severity for the identified disease.

3.1 Image Preprocessing

All leaf images used in this work are downloaded from reliable web sites for research on plant diseases. Domain experts from the Ethiopian Institute of Agricultural Research (EIAR) examined the validity of the downloaded images. However, each image might be captured from different distance in different environmental settings. So image preprocessing is needed to enhance the images. To come up with better results, we applied image resizing and image adjustment techniques.

We downloaded a total of 66 images which is not enough to train the model. To increase the effectiveness of the model, some techniques are applied to increase the data size like altering noises on the original image where 'Salt and pepper' noise is added. In addition to adding noise, flip technique is also applied to the images. As a result, a total of 3* 66 = 198 images are used as a dataset.

3.2 Image Segmentation

Before segmenting the images, color space conversion is done from RGB to L*a*b*. This conversion is required because the L*a*b* color space is device independent which means the device from which it is created or it is displayed does not affect its color definition [15, 16, 17]. In addition, this color space is more preferable in representing color [15, 16]. In L*a*b* color space, all the color information exists in a* and b* spaces while the L represents the luminosity [15, 18]. It also best fits the segmentation technique we use which is K-means.

Image segmentation is the most important part in image processing. It is performed to divide an entire image into several parts. This division is made simplify image analysis. Segmentation may depend on various features of the image like color and texture. In this paper, the segmentation technique preferred to be used is the color based k-means clustering. K-means is a partitioning technique that clusters a given data X into k mutually exclusive clusters [19]. Then it assigns an index of the cluster (idx) for each observation using Equation 1.

$$idx = kmeans(X, k) \quad (1)$$

K-means starts with assigning a centroid for each cluster. The number of clusters is pre-defined and initialization of the center of each cluster is done by squared Euclidean distance metric and the k-means++ algorithm [21, 22].

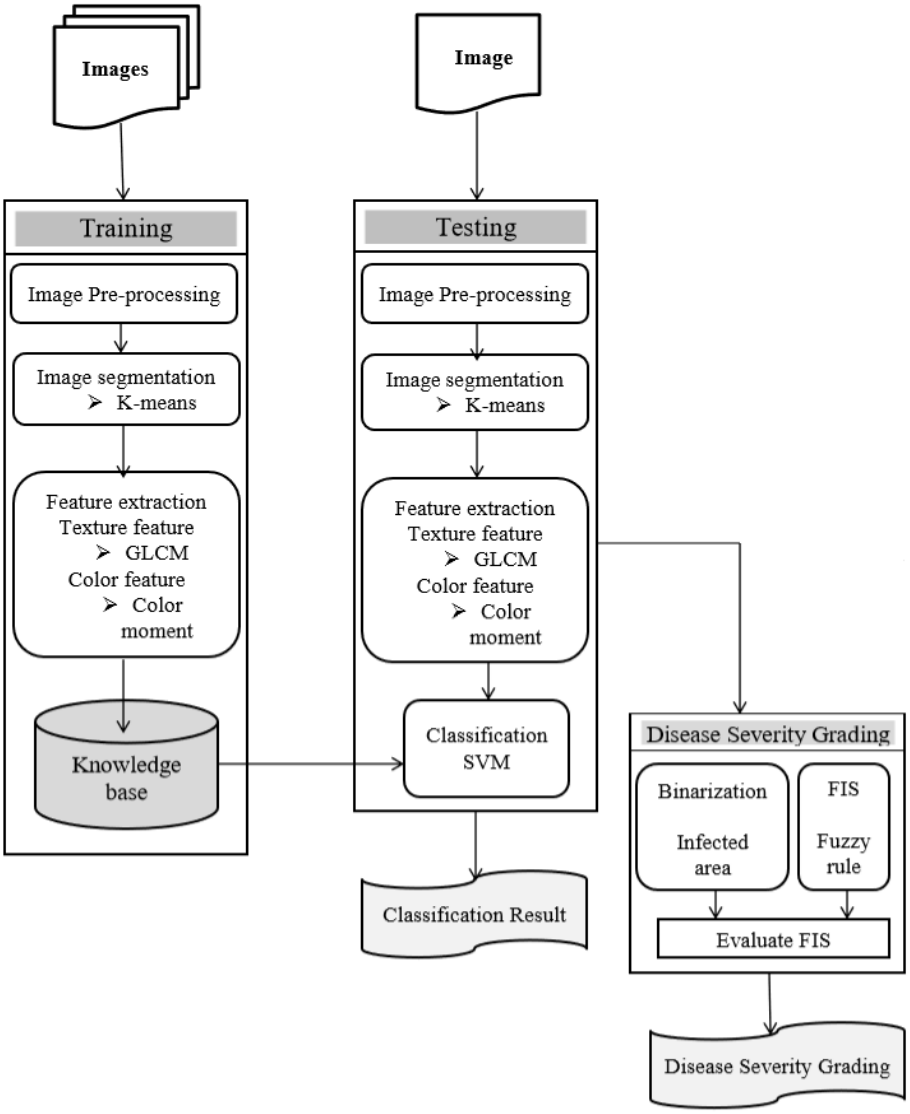


Figure 1: The Proposed System Architecture of Oilseed Crop Leaf Disease Identification

3.3 Feature Extraction

Feature extraction is required when the input image is assumed to be too large to be processed and notoriously (much data, but not much information) redundant. There are different feature extraction methods. In this paper color and texture feature extractions are considered.

a. Color Feature Extraction

Color moment is used for extracting color features. In color moments, images are assumed to be represented by a probability distribution of colors [36]. Keen [36] defines moments as the characterizations of those probability distributions. As Stricker and Orengo [24] stated, if the probability distributions of images follow certain pattern, those images are assumed to be similar. So the measured moment of the image can be used for extracting color features of an image [36]. The first three moments *Mean*, *Standard Deviation*, and *Skewness* are used by Stricker and Orengo [24]. We used Mean and Standard Deviation as feature variables.

The first color moment, Mean, can be stated as the average color value in the image whereas the second color moment, Standard Deviation, is the square root of the Variance of the distribution [36].

b. Texture Feature Extraction

There are several techniques in image processing for extracting texture features. In this work, GLCM is used.

GLCM uses graycomatrix function. This function calculates GLCM by considering spatial relationship of pixels (those pixels that are in gray-level) [30, 31, 32] as shown in Figure 2 [33]. Each element (i, j) in the resultant GLCM is simply the sum of the number

of times that the pixel with value i occurred in the specified spatial relationship to a pixel with value j in the input image.

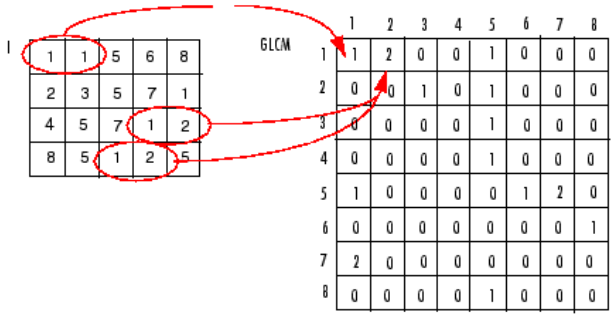


Figure 2: Process of Creating GLCM

Statistics can characterize the texture of an image because they provide information about the local variability of the intensity values of pixels in an image [33]. Statistical measures such as local standard deviation of an image and local entropy of a grayscale image are obtained from the created GLCM. The texture features are extracted using these statistical measures. In this paper using GLCM, 22 texture features are extracted. They are used for training the SVM. The selection of those texture features is based on the accuracy result they have brought during training of the model.

3.4 Image Classification and Identification

The extracted features are used as input for this step. In this step machine learning approaches are followed in many researches. Among those the supervised learning approach multi-SVM with ECOC (Error-Correcting Output Coding) algorithm is used in this paper. SVM is known for its suitability for classification when there exist few number of datasets or many attributes are considered [11, 29]. This is the major reason that this approach is selected.

Since SVM uses different kernel functions, for training of the model comparison of accuracy rate is done among kernel functions. The cubic kernel function is selected for training the model, which has provided better accuracy. This model then predicts the category/class for new data based on the dataset that it is trained with. Hence, it identifies the type of leaf disease that the plant is infected with.

3.5 Disease Severity Grading

Other than identifying oilseed crop leaf diseases, the proposed solution also provides a way for knowing in which level (stage) the disease is. For doing so fuzzy logic is applied.

In this paper oilseed severity grading Fuzzy Inference System (FIS) is designed for grading the stage of the diseases. For doing this, the percentage of area of infected region of the leaf is taken as an input for the system and the stage of the severity is given as an output.

For calculating the percentage of infected region, first both the segmented and the whole images are binarized. Then the sum of white pixels is segmented and binrazed image is divided by the sum of pixels in the whole and binarized images. Finally, the result is multiplied by hundred to get the percentage of the infected region of the leaf image.

For both input (infected area) and output (severity stage), a triangular membership function is used since it is the simplest and the most effective [34]. Triangular membership function can be represented mathematically as [35]:

$$f(x,a,b,c) = \max\left\{min\left(\frac{x-a}{b-a}, \frac{c-x}{c-b}\right), 0\right\}$$

where a and c are the “feet” of the triangle and b is the “peak”.

In our prototype the severity has six stages. The FIS predicts the stage of severity (output) based on fuzzy rules that have already been set during the design of the FIS.

4. Result

Classification learner app of Matlab is used for training and testing the model. Multi-SVM is chosen for modelling the prototype. Among quadratic, Gaussian and cubic kernel functions that the SVM uses, best accuracy is achieved while training the model with cubic kernel function.

The dataset prepared in the preceding image processing phases are used to evaluate the model. A total of 28 features which consist of 22 texture features and 6 color features are used. The total dataset consists of 198 datasets. The dataset consists of both diseased and healthy leaves of oilseed (soybean and sunflower). All these data are divided

for training (training set) and testing (testing set). During training 162 (82% of the dataset) of the data are used for both training and validating and the remaining 36 test set (18% of the dataset) are used for testing. During the test almost all the test data are correctly classified to their category or class. Accuracy of 97.2% is achieved with overall error of 2.8%. Accuracy is calculated as the sum of correct classification (true positive) divided by the total number of classifications.

5. Conclusion and Future Works

With the proposed solution a very good result is achieved for identification of oilseed leaf diseases as well as disease severity grading. This will make a big difference in the experience and trends of oilseed leaf disease identification methods which is manual in most developing countries. The system has impact on reducing the time it takes and the effort required for identifying diseases. Since domain experts are few, especially in developing countries, it supports domain experts as well as farmers who may not have adequate knowledge on identifying diseases. The treatment and measures taken after a crop is identified as infected depends on the level of disease severity. On top of the identification, the proposed solution has also the capability of leveling disease severity which helps to mitigate diseases in appropriate way. This boosts the efficiency of disease identification and management experiences. This in tern prevents bigger yield loss. So it plays a significant role in sustaining the production of oilseed crops.

Severity grading of oilseed diseases works well by considering the percentage of affected regions. Future work will consider the development stage of the plant that is believed to provide better results.

References

[1] World Bank, “Agriculture and Food”, https://www.worldbank.org/en/topic/agriculture/overview, Last accessed on September 12, 2019.

[2] Neha Khanna and Praveen Solanki, “Role of agriculture in the global economy,” in Proceedings of the 2nd International Conference on Agricultural and Horticultural Sciences, Hyderabad, India, February 2014.

[3] United States Department of Agriculture, “Global Markets: Oilseeds – Consumption Grows Despite Slowing Trade, Production, “2019, https://agfax.com/2019/06/12/global-markets-oilseeds-consumption-grows-despite-slowing-trade-production, Last accessed on December 06, 2019.

[4] European Commission, “Oilseeds and Protein Crops market situation,” https://circabc.europa.eu/sd/a/215a681a-5f50-4a4b-a953-e8fc6336819c/oilseeds-market%20situation.pdf, Last accessed on December 01, 2019.

[5] Biruk Ambachew, “Knowledge Based System for Oilseed Crop Disease Diagnosis and Treatment,” Addis Ababa University, March 2015.

[6] Sachin D. Khirade and A. B. Patil,” Plant disease detection Using image processing,” in Proceedings of the IEEE International conference on computing communication control and automation, February 2015.

[7] Surender Kumar and Rupinder Kaur, “Plant Disease Detection using Image Processing - A Review”, International Journal of Computer Applications, Vol. 124, No. 16, 2015.

[8] Pallavi S. Marathe, “Plant Disease Detection using Digital Image Processing and GSM,” International Journal of Engineering Science and Computing, Vol. 7, No. 4, 2017.

[9] Jiangsheng Gui and Mor Mbaye, “Identification of Soybean Leaf Spot Diseases using Deep Convolutional Neural Networks”, International Journal of Engineering Research & Technology, Vol. 08, No. 10, 2019.

[10] F. Jakjoud, A. Hatim, and A. Bouaddi, “Detection of diseases on tomato leaves based on Sub-Classifiers Fuzzy Combination”, International Journal of Innovative Technology and Exploring Engineering (IJITEE), Vol. 8, No. 5, 2019.

[11] Sonal P. Patil and Rupali S. Zambre, ”Classification of Cotton Leaf Spot Disease Using Support Vector", International Journal of Engineering Research and Applications, IJERA, Vol. 4, No. 5, pp. 92-97, May 2014.

[12] Vinaya Mahajan and N. R. Dhumale, “Leaf Disease Detection Using Fuzzy Logic”, International Journal of Innovative Research in

Science, Engineering and Technology (IJIRSET), Vol. 7, No. 6, 2018.

[13] Juhi Singh, "A Parameterized Comparison of Fuzzy Logic, Neural Network and Neuro Fuzzy System: A Literature“, International Journal of Computer Science and Mobile Computing, IJCSMC, Vol. 5, No. 5, 2016.

[14] Mahesh Prasanna and Shantharama Rai, "Applications of Fuzzy Logic in Image Processing – A Brief Study”, International Journal of Advanced Computer Technology, Vol. 4, No. 3, 2015.

[15] Mokrzycki W. S. and Tatol M., ”Perceptual difference in L*a*b* color space as the base for object colour identfication”, in Proceedings of the 1st International Conference on Image Processing and Communications, 2014.

[16] Vivek Singh Rathore1, Messala Sudhir Kumar, and Ashwini Verma, “Colour Based Image Segmentation Using L*a*b* Colour Space Based on Genetic Algorithm”, International Journal of Emerging Technology and Advanced Engineering, IJETAE, Vol. 2, No. 6, 2012.

[17] Dibya Jyoti Bora, “Importance of Image Enhancement Techniques in Color Image Segmentation: A Comprehensive and Comparative Study”, Indian Journal of Scientific Research, Vol. 15, pp. 115-131, 2017.

[18] Richard S. Hunter, “Photoelectric Color Difference Meter”, Journal of The Optical Society of America, Vol. 48, No. 12, pp. 985-995, 1958.

[19] Aastha Joshi and Rajneet Kaur, ”A Review: Comparative Study of Various Clustering Techniques in Data Mining”, International Journal of Advanced Research in Computer Science and Software Engineering, IJARCSSE, Vol. 3, No. 3, 2013.

[20] MathWorks, “k-Means and k-Medoids Clustering,” Statistics and Machine Learning Toolbox, 2015.

[21] David Arthur and Sergei Vassilvitskii, ”k-means++: The Advantages of Careful Seeding”, in Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA, New Orleans, Louisiana, 2007.

[22] Youguo Li and Haiyan Wu, “A Clustering Method Based on K-Means Algorithm”, in Proceedings of the 2012 International Conference on Solid State Devices and Materials Science, Macao, 2012.

[23] Abdallah A. Alshennawy and Ayman A. Aly, “Edge Detection in Digital Images Using Fuzzy Logic Technique”, International Journal of Computer and Information Engineering, Vol. 3, No. 3, 2009.

[24] Markus Stricker and Markus Orengo, "Similarity of Color Images”, in Proceedings of SPIE - The International Society for Optical Engineering, San Jose, 1995.

[25] Surender Kumar and Rupinder Kaur, “Plant Disease Detection using Image Processing - A Review”, International Journal of Computer Applications, Vol. 124, No. 16, 2015.

[26] S. Lokesh, D. Naveenkumar, K. Rajesh, G. A. Ramesh Kamath, and Martin Joel Rathnam, “Leaf Disease Detection and Grading using Computer Vision Technology and Fuzzy Logic”, International Journal of Innovative Research in Science, Engineering and Technology, IJIRSET, Vol. 6, No. 3, 2017.

[27] Chaitali G. Dhaware and K. H. Wanjale, "A Modern Approach for Plant Leaf Disease Classification which Depends on Leaf Image Processing", in Proceedings of the International Conference on Computer Communication and Informatics (ICCCI), Coimbatore, India, January 2017.

[28] Jun Liu, Fang Lv, and Bobo Liu, "Identification Method of Sunflower Leaf Disease Based on SIFT Point", Journal of Image and Graphics, Vol. 7, No. 2, pp. 64-67, 2019.

[29] Jain S., “A Machine Learning Approach: SVM for Image Classification in CBIR”, International Journal of Application or Innovation in Engineering and Management, Vol. 2, No. 4, 2013.

[30] P. Mohanaiah, P. Sathyanarayana, and L. GuruKumar, “Image Texture Feature Extraction Using GLCM Approach”, International Journal of Scientific and Research Publications, Vol. 3, No. 5, 2013.

[31] Leen-Kiat Soh and Costas Tsatsoulis, “Texture Analysis of SAR Sea Ice Imagery Using Gray Level Co-Occurrence Matrices”, Journal of IEEE Transactions on Geoscience and Remote Sensing (TGRS), Vol. 37, pp. 780-795, 1999.

[32] Robert M. Haralike, K. Shanmugam, and Itshak Dinstein, "Textural features for image classification", Journal of IEEE Transactions on Systems, Man, and Cybernetics, Vol. 3, No. 6, 1973.

[33] MathWorks, "Texture Analysis", Image Processing Toolbox, 2015.

[34] Jin Zhao and Bimal K. Bose, "Evaluation of membership functions for fuzzy logic controlled induction motor drive", in Proceedings of the IEEE 2002 28th Annual Conference of the Industrial Electronics Society, IECON 02, Sevilla, Spain, 2002.

[35] MathWorks, "Fuzzy Inference System Modeling", Fuzzy Logic Toolbox, 2015.

[36] Noah Keen, "Color Moments," 2005, http://homepages.inf.ed.ac.uk/rbf/CVonline/LOCAL_COPIES/AV0405/KEEN/av_as2_nkeen.pdf, Last accessed on December 02, 2019.

Tigrigna Text to Ethiopian Sign Language Machine Translation

Redwan Tahir
HiLCoE, Computer Science Programme, Ethiopia
redtah2007@yahoo.com

Michael Melese
HiLCoE, Ethiopia
michael.melese@aau.edu.et

Abstract

Language is a method of human communication, either spoken, written or symbol, using words in a structured and conventional way. There are roughly 7,111 spoken languages in the world today, which include Sign language for the hearing-impaired. Even though there are a huge number of deaf people in Tigray, there is still marginalization of the deaf community regardless of their will and capacity because of communication gap. ICT can play a role in bridging this gap.

The objective of this paper is to design and develop Tigrigna text to Ethiopian Sign Language (EthSL) machine translation system. We studied the linguistic structure of both languages and learned EthSL to maintain Tigrigna words dictionary containing root words and their morphemes, which help to design Tigrigna text to EthSL translator and develop a prototype. The methodology applied is Design Science using Dictionary based machine translation and different Natural Language Processing (NLP) tools were utilized.

The prototype consists of modules for text analysis, mapping and sign animation (using 3D avatar). The prototype is designed as a web application. Using the translator, Tigrigna texts can be translated into gestures, which can be understood by hearing-impaired people. The hearing-impaired can also utilize the software to develop written language skills. Moreover, it may be used as a teaching tool for Ethiopian sign language. The prototype was evaluated in terms of capability of signing. The evaluation result shows an accuracy of 80% for words, 85% for letters, and 75% for numbers.

Keywords: Hearing-Impaired; Deaf; Sign Language; Tigrigna Language; ICT; 3D Avatar; EthSL

1. Introduction

Language is a method of human communication, either spoken, written or symbol, using words in a structured and conventional way [1]. There are roughly 7,111 spoken languages in the world today of which 30% are in Africa and 85 are in Ethiopia, one of them being Tigrigna [2]. It is very difficult to understand every language but a language is the only medium of communication for humans. The idea of translation is to communicate messages from one language to others.

Our aim is to help hearing-impaired people communicate with hearing persons who speak Tigrigna using sign language by developing a software system.

a. Tigrigna Language

Tigrigna is spoken in the East African countries of Ethiopia and Eritrea. It is one of the two official languages of the State of Eritrea. It is also a working language of the Tigray region of Ethiopia [3]. According to the 2007 census, it is spoken by 4,467,000 people in Ethiopia, of which 4,320,000 are First Language (L1) speakers and 147,000 are Second Language (L2) speakers. There are 2,820,000 monolinguals.

Tigrigna is a Semitic language of the Afro-Asiatic family that originated from the ancient Geez language. It is closely related to Amharic and Tigre. Tigrigna is written in Geez script, also called Ethiopic [3]. Ethiopic script is syllabic, i.e., each symbol represents 'consonant + vowel' combinations. Each Tigrigna base character, also known as Fidel, has seven vowel combinations. Tigrigna is written from left to right. It is a 'low resource' language having very limited resources: data, linguistic materials and tools [3].

b. Sign Language (SL)

Sign language is a visual gesture language. It is used by hearing-impaired people for communication, which is done by using hands, face, arms and body. Sign language is not universal. Every country has its own sign language which differs from other countries in syntax and grammar [4].

Ethiopia also has a sign language as other known sign languages which is known by the name of Ethiopian Sign Language (EthSL) and its typology is in Amharic script one handed finger spelling. According to the Ethiopian Association report of 2005 [5], 1,000,000 of the population of Ethiopia is deaf with unknown number of isolated deaf in rural areas.

Ethiopian sign language has been used in various Ethiopian schools for the deaf since 1971 and at the primary level since 1956. Presumably a national standard, it is used in primary and secondary schools, at Addis Ababa University, and on national television. The Ethiopian deaf community uses the language as a marker of identity [6].

c. Machine Translation (MT)

Machine translation refers to the translation of text from one language (source language) to another (target language) without the help of a human. The aim of machine translation is to provide a system that translates text of source language to target language with the same meaning as in source language [7].

The translation has two steps: *decoding* the source text and *encoding* it into the target language. Both decoding and encoding require deep knowledge of the source and target languages. This includes grammar, semantics, and syntax understanding of both languages. Natural language is very complex in terms of word meaning and grammar rules [7].

In Ethiopia, there is still marginalization of the deaf community. Regardless of their will and capacity, their contribution to the country's development is minimal [8]. In order to ease the communication problems of the deaf and HoH in their day-to-day activities, currently human translators play an important role. Therefore, it is imperative to fill the communication gap using appropriate tools and assistive technology.

Dagnachew Feleke [9] developed an ESL synthesizer of Amharic text at world level and Lily Abebe [10] at sentence level.

Minilik Tesfaye [11] used a rule based approach called Interlingua machine translation. However, the study has two major gaps. First, the prototype outputs the text spelling in sign language not in meaning so the deaf must know what the combination of letters means in Amharic. For example, if we say ቡላ the system reads the word ቡ and ላ so the user must know what ቡላ means. Second, the author used Interlingua approach which means that there is a mediator independent language so it decreases its speed and accuracy.

In conclusion, all the studies focused on Amharic and can't be directly applied to Tigrigna since the two languages are morphologically different. There are also gender differences in some words. Hence, this study tries to address the gap using Rule based MT to translate Tigrigna text to EthSL.

2. Related Work

A system was proposed in [14] to translate English to sign language using a set of hand-coded schemata as an interlingua. While the implementation focused on American Sign Language, the framework also considered British, Irish, and Japanese Sign Languages as well. The study used Artificial Intelligence (AI) knowledge representation, metaphorical reasoning, and blackboard system architecture.

During the analysis stage, English text undergoes sophisticated idiomatic concept decomposition before syntactic parsing in order to fill slots of particular concept/event/situation schemata.

The advantage of the logical propositions and labeled slots provided by a schemata-architecture is that common sense and other reasoning components in the system could later easily operate on the semantic information. Because of the amount of hand coding needed to produce new schemata, the system would be feasible only for limited domains. To compensate for these limitations, if a schema did not exist for a particular input text, the researchers suggested that the system could instead perform word-for-sign transliteration (producing Signed English output).

Another study on this area was TEAM [15]. It is an English-to-ASL translation system tree Adjoining Grammar rules to build an ASL syntactic structure simultaneous to the parsing of an English input text and the output of the linguistic portion of the system was a string-like ASL "gloss notation with embedded parameters" that encoded limited information about

morphological variations, facial expressions, and sentence mood as features associated with the individual words in the string.

Minilik Tesfaye [11] developed an avatar-based translation system by adopting a machine translation approach that can take Amharic text as an input and generates an EthSL equivalent of the Ethiopian Finger alphabet input text. Its main drawback is that it uses finger spellings instead of using the corresponding signs of the word in EthSL to spell the Amharic word. However, finger spelling is a manual code to represent the letters of the spoken language alphabet and is mostly used to illustrate words that have no corresponding sign. Since the researcher used only Finger spelling in the translation, the translation from Amharic to EthSL cannot be taken as a translation.

Another study was by Abadi Tsegay [13] which focused on automatic translation of Amharic text to Ethiopian sign language with personal and possessive pronouns as they are bound morphemes on verbs and nouns. The system architecture of the proposed system is composed of three modules: input text analysis, text to sign mapping and sign synthesis/generation.

Dagnachew Feleke [9] also studied MT system to translate Amharic words to their equivalent representation in EthSL. The system uses rule-based machine translation (RBMT) to map a given Amharic word to EthSL equivalent word using morphological analysis and synthesis. The scope is limited to morphological level translation, i.e., direct RBMT architecture was followed and it accepts Amharic words in text format as input and outputs the equivalent in EthSL by using signing avatar. The HamNoSys and SiGML tools were applied for generating EthSL animation in order to play in the avatar, and fifty randomly selected words and all alphabets were used to test its performance.

The system architecture of the proposed system has three major components and one supporting system. The first is Amharic text processing to change a given Unicode written Amharic input to ASCII using the SERA representation. The second one is Morphological Analyzer that contains the morphological expander, maintains the Amharic dictionary and when invoked, it will load the dictionary to memory with all the needed information. The third one is Amharic-EthSL mapping in order to change a given text in Amharic to EthSL and the supporting part is Morphological expander which is a morphological analyzer preparing the words and holds the conversion key.

Amharic Sentence to Ethiopian sign language was developed by Daniel Zegeye [12] using RBMT. It supports simple Amharic phrases, letters and numbers, and outputs EthSL 3D animation with system accuracy of 58.77%, 75.76%, and 84%, respectively, at phrase, letter, and number levels. Its main strength is that it considers the translation of Amharic text into EthSL at sentence level.

3. The Proposed Solution

The proposed prototype's main goal is to help the deaf community and to fill the communication gap between hearing and deaf communities. The overall implemented process starts from problem identification till the last called communication in order to develop the prototype.

After extensive study on the subject, the system is designed and the implementation for small vocabulary is done. This section presents a brief explanation of each stage of the System Architecture and prototype design and its components. The overall process has 3 main components as shown in Figure 1:

- Interface,
- Backend, and
- Knowledge Database.

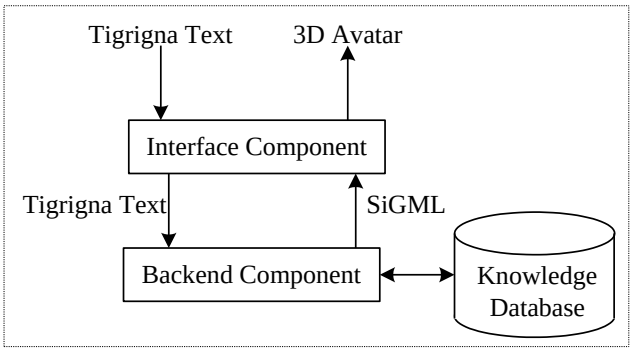


Figure 1: General Architecture of the Developed System

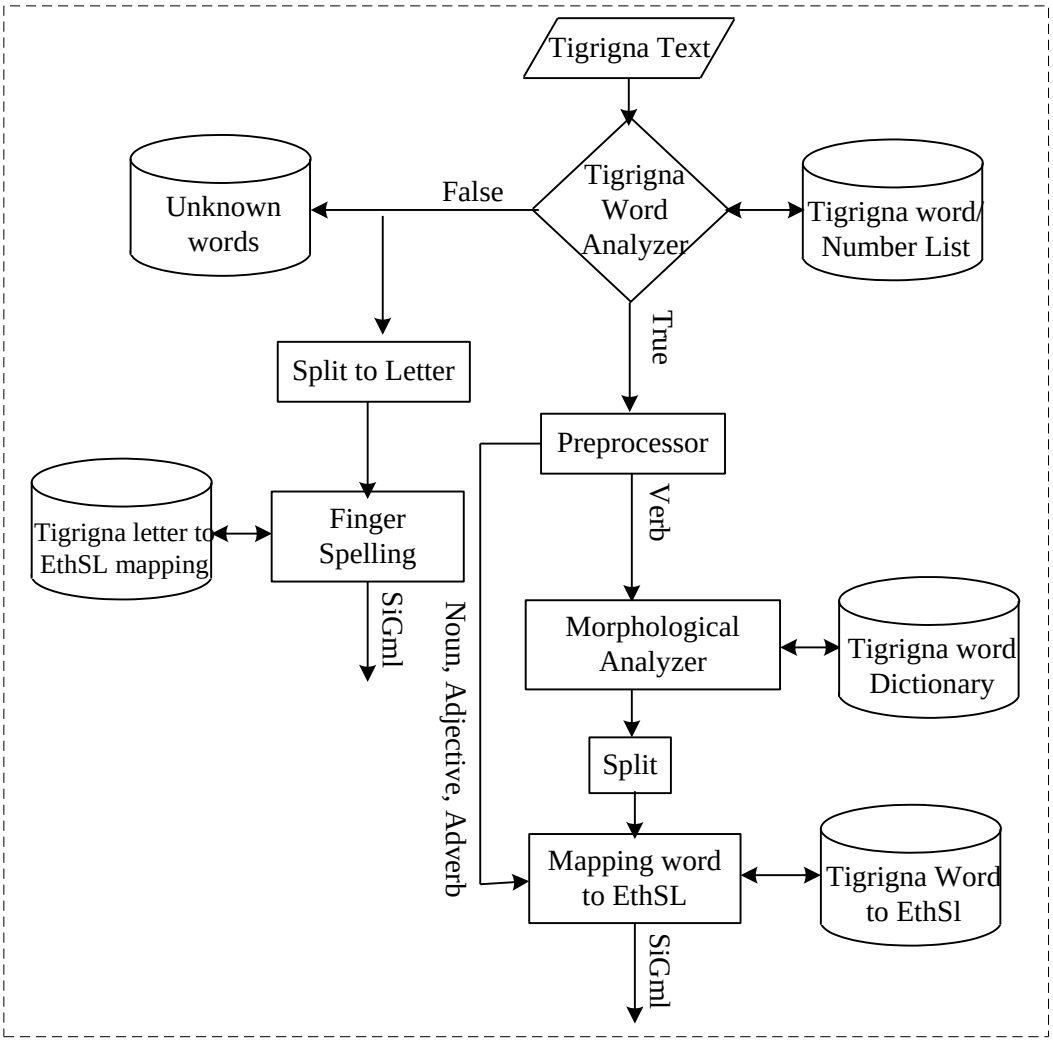


Figure 2: Architecture of the Developed Prototype

1. Interface

The *Interface* is the main component that serves as intermediary between the user and the *Backend* component and it has two pats. The first is the Web interface that accepts a Tigrigna word from the user as input and passes it to the *Backend* component for further processing. The second one accepts SiGML record scripts from the *Backend* component and plays the script in 3D Avatar animation for the user which can easily be understood.

3D avatar is a gesture representation by means of VGuido (the eSIGN 3D avatar) animations that seems human-like motion. It is a plug-in for browsers. In order to make an avatar sign or gesture, pre-specified animation sequences must be sent to the avatar.

In addition, those sequences must be consistent with temporal sequence of frames, each of which defines a static posture of the avatar at the appropriate moment and generated synthetically from an input script in the SiGML notation.

2. Backend Component

The *Backend* component accepts a Tigrigna word from the *Frontend* component and outputs SiGML files. It has the following three modules.

- Tigrigna Text Analysis,
- Natural Language Processing (NLP), and
- Text-to-sign mapping.

a. Tigrigna Text Analysis

The task of this module is to accept a Tigrigna word and map it to Tigrigna word list corpus. It checks the word in the knowledge database and if the word exists, it passes it to the NLP module. If the word does not exist in the knowledge database, it will play in finger spelling and register the word in *Unknown words* and the admin will later include it in the Tigrigna word list if it is a Tigrigna word or a number.

b. Natural Language Processing (NLP)

The NLP module handles all natural language processing tasks needed by the system. This module has the following three sub-components.

- POS tagger,
- Preprocessor, and
- Morphological analyzer.

i. Preprocessor

The preprocessor checks the return POS of the word and passes to the next stage which has 3 possibilities. If the POS tag is a verb, it will pass the word to the Morphological analyzer. If it is a proposition it will terminate the process by displaying a message for the user because there is no preposition in sign language. If the word is out of both options, it will pass the word to the text- to-ESL map module.

ii. Morphological Analyzer

This is the backbone of the MT system. It contains the morphological expander that expands a Tigrigna verb into separate writing way from Tigrigna dictionary that was developed for this purpose. The dictionary of each word contains information about the stem word and object or subject of the verb. The generated verb is used as a key in the dictionary and the value is the verb (word) with its object or subject information attached to it.

c. Text-to-Sign Mapping Module

The Text-to-sign mapping module is responsible for mapping of text to SiGML file sign script. This module accepts a Tigrigna word in EthSL format and finds the word from text to SiGML mapping knowledge database and sends to the interface the path of the SiGML source. The word is used as a key in the dictionary and the value is the SiGML file path.

3. Knowledge Database

The knowledge database works by checking if a word or a number is Tigrigna or not and if the word is found and its POS tag is verb, then it splits the word to extend it morphologically. But if the POS tag is not a verb, it will bypass the morphological extender. If the word is not found, it will sign using finger spelling. Finally, it aligns the word or number with its equivalent meaning of Ethiopian sign language word or finger spelling.

4. Conclusion and Future Work

In this paper we designed and developed a prototype to be used to translate Tigrigna text into EthSL. Written text interpretation could be invaluable to members of the hearing-impaired community, especially in areas of low availability of human interpreters.

Animation of sign language via an avatar is a potentially promising method of presenting sign language information. To present EthSL signs, Tigrigna text-to-sign translator uses 3D avatar animation, which is cost effective and flexible for signing sign languages as method of animation.

EthSL is inclusive of both signing and finger spelling. Signing includes the expressions of the conceptual sign that are mainly used to convey meaning in EthSL.

In order to develop the sign language for each of the words, first we translated Tigrigna words into English using dictionary and then we entered those English words to HamNoSys and got the HamNoSys notation of the word. After that we entered this notation in to eSign editor and received the Sigml file in Xml format equivalent to the word that was entered. Finally, the 3D avatar was played and checked by EthSL interpreter whether it is correct or not. When there is an incorrect interpretation, it is manually edited in the HamNoSys tool.

The system provides a one-way interface for communication by translating Tigrigna text into EthSL. Nonetheless, the ideal sign language interface would be one that accepts Tigrigna text as input while also being able to display it in EthSL.

Future works include the following:

- Building Tigrigna morphological analyzer, especially for sign language synthesis considering critical features.
- At present, the prototype provides translation of words, letters and numbers in Tigrigna. To provide translation services for Tigrigna phrases, sentences and documents, the system needs to be extended.

- Building a larger size parallel corpora of Tigrigna to EthSL sentences.
- To support a wider range of Tigrigna input, the underlying dictionary of the Tigrigna text to EthSL translator needs to be enriched and extended.
- In order to allow a complete conversation between a hearing person and a deaf person (in both directions), it is important not only to translate Tigrigna text into sign language, but also to produce Tigrigna text from sign language.

References

[1] Lexico power by Oxford definition of Language, https://www.lexico.com/en/definition/language, Last accessed on 02 Dec. 2019.

[2] Ethnologue: Languages of the World, https://www.ethnologue.com/, Last accessed on 02 Dec. 2019.

[3] Omer Osman Ibrahim and Yoshiki Mikami, "Stemming Tigrinya Words for Information Retrieval", December 2012.

[4] Khushdeep Kaur and Parteek Kumar, "HamNoSys to SiGML Conversion System for Sign Langu", Thapar University, India, June 2012.

[5] Lewis, M. Paul, Gary F. Simons, and Charles D. Fennig (eds.), "Ethnologue: Languages of the World", Eighteenth edition. Dallas, Texas: SIL International, Online version, http://www.ethnologue.com.

[6] Wikipedia, "Ethiopian sign languages", https://en.wikipedia.org/wiki/Ethiopian_sign_languages, Last accessed on 02 Dec. 2019.

[7] Muhammad Irfan, "Machine Translation", Department of Computer Science, Bahria University, Islamabad, October 2017.

[8] K. Kaur and P. Kumar, "HamNoSys to SiGML Conversion System for Sign Language Automation", Thapar University, India, June 2012.

[9] Dagnachew Feleke, "Machine Translation System for Amharic Text to Ethiopian Sign Language, Unpublished Masters Thesis, Addis Ababa University, October 2011.

[10] Lily Abebe, "Example Based Amharic Text to Ethiopian Sign Language Machine Translation", Unpublished Masters Thesis, Addis Ababa University, June 2018.

[11] Minilik Tesfaye, "Machine Translation Approach to Translate Amharic Text to Ethiopian Sign Language", Unpublished Masters Thesis, St. Mary's University College, Addis Ababa.

[12] Daniel Zegeye, "Amharic Sentence to Ethiopian Sign Language Translation", Unpublished Masters Thesis, Addis Ababa University, 2015.

[13] Abadi Tsegay, "Offline candidate hand gesture selection and trajectory determination for continuous Ethiopian sign language", Unpublished Masters Thesis, Addis Ababa University, 2011.

[14] Tony Veale and Alan Conway, "Cross Modal Comprehension in ZARDOZ An English to Sign-Language Translation System, Hitachi Dublin Laboratory, O'Reilly Institute, Trinity College, Dublin, Ireland, 1994.

[15] Achraf Othman and Mohamed Jemni, "Statistical Sign Language Machine Translation: from English written text to American Sign Language Gloss", Research Lab, UTIC, University of Tunis, September, 2011.

Near Real-time SIM Box Fraud Detection using Stream Mining

Sisay Matebe
HiLCoE, Software Engineering, Ethiopia
EthiopiaSisaymatebe@gmail.com

Mesfin Kifle
HiLCoE, Ethiopia
Department of Computer Science, Addis Ababa University, Ethiopia
kiflemestir95@gmail.com

Abstract

Voice traffic termination fraud, often referred to as Subscriber Identity Module box (SIM Box) fraud, is a common illegal practice on mobile operators. It terminates international calls using a local SIM card inside the SIM Box machine and delivered to the customer as local call without operator's know how to manipulate the tariff difference. As a result, operators are losing billions of money annually. In this paper, we propose near real-time SIM Box fraud detection using stream mining analytics technique to enable the system real-time decision making by inspecting, correlating and analysing the telecom data as it is streaming into applications. Rather than singling out specific transactions, our solution analyses historical transaction data to model a system that can detect fraudulent patterns. This model is then used to analyse Call Detail Records (CDR) transactions in real-time.

We have employed Lammed architecture with seven time-sensitive data mining classification algorithms, namely Decision Tree, Extra Trees Classifier (random), Logistic Regression and Random Forest, Linear SVM (Support Vector Machine), MLP Neural Network, and Gradient Boosting Classifier. Six selected attributes from the CDR that could be used to detect fraudulent behaviour in a near-real time are extracted. The experiment set up will enable us to understand SIM Box from genuine call behaviour both in near-real-time and one-hour window (eleven attributes selected), in case near-real-time detection missed to detect. Finally, we obtained a good result from Very Fast Decision Tree (VFDT) classification for near-real time and linear SVM for one-hour detection. VFDT classification algorithm achieved an accuracy of 98.5% and false positive 0.4% with average latency of 2 minutes per detection at input event rate of 280 CDR per second with limited hardware. SVM classification achieved 99% accuracy with 0.2% false positive for one-hour window detection. Finally, we propose both near-real-time and one-hour SIM Box fraud detection solution for effective outcomes.

Keywords: Stream Event Processing; Stream Mining; Data Analytics; Call Detail Records (CDR); Call Patterns; SIM Box Fraud; Telecom Fraud

1. Introduction

Telecommunication Fraud can be defined as any activity by which service is obtained with intention of not to pay [2]. Imagine you get an unknown call from a local number, you pick up the phone and it is from a friend or a family living abroad, it does feel strange that you would receive an international call from a local number; this is basically the SIM Box fraud, or SIM Boxing. In this case, the international calls are transferred over the Internet to inject them back into the cellular network through SIM Box machine with multiple local SIM cards inside. As a result, the calls turn as local to the destination network and the fraudsters who set up these boxes pay only local rates for mobile operators after charging international rates from the source operators.

Currently available traditional fraud detection systems make detections by generating a set of features or aggregate values by querying static data over a large time window and make decisions based on such values [3, 4]. This is time consuming and by that time a fraudster can make a number of successful calls before being detected. SIM Box Fraud has a big business impact to telecom operators and demands near real time CDR analysis tool which can watch CDR streams and generate a recent or near-real time view of the incoming CDR. Such approaches support the operators to gain maximum advantage by exploiting timeliness of detected events

and save significant amount of revenue by minimizing fraudulent activities.

In this paper, a real time detection system is designed and tested in cooperation with ethio telecom. The system utilizes stream mining techniques to detect whether a SIM card is used by a normal customer or by a SIM box fraudster.

2. Literature Review

2.1 Telecom Fraud Types

The telecom industry is one of the major sectors in the world infiltrated by fraud [1, 2]. There are different types of telecommunication frauds and these can occur at various levels. Different authors categorized frauds in different ways. For instance, the work in [8] categorized subscription fraud and technical fraud in general as the most common fraud types, whereas the work in [4] classifies them into seven groups as Superimposed, Subscription, Technical, Internal Fraud, Social Engineering, Fraud based on loopholes in technology, and Fraud based on new technology. Similarly, the work in [5] lists the top three types of telecommunication frauds that cause a significant loss to the operators as International Revenue Share Fraud, Premium Rate Service Fraud and Bypass fraud. In [14], they are classified as Contractual, Hacking, technical and Procedural frauds.

a. International Revenue Share Fraud (IRSF)

IRSF is the largest contributor to the overall fraud losses according to Communication Fraud Control Association (CFCA). It is when a fraudster makes an agreement with a local carrier in high cost destination to share profit for increasing traffic, then the fraudster hacks into any organization's public branch exchange (PBX) and gets illegal access to generate calls. After that, the fraudster generates high traffic calls to high cost destinations and gets revenue from the sharing agreements [1].

b. Premium Rate Service Fraud

Premium rate service is an agreement between some service provider and telecom companies to share revenues generated by traffic to the premium service number. It is used in TV shows, contests and entertainment services. The fraudsters try to stimulate consumers by giving them a missed call or a message, then make money from share of the call back revenues [4].

c. SIM Box Fraud

Interconnect Bypass (SIM Box) fraud is unauthorized insertion of traffic onto another carrier's network. This fraud type can be named as Interconnect fraud, Global System for Mobile Communications (GSM) Gateway fraud, or SIM Boxing. This scenario requires that the fraudsters have access to advanced technology such as VoIP, which can make international calls appear to be cheaper domestic calls, effectively bypassing the normal payment system for international calls. The fraudsters will typically sell long distance calling cards. When customers call the number on the cards, operators can switch the call to make it look like a domestic call. The most common implementation of interconnect bypass fraud is known as SIM Boxing, which is enabled by VoIP GSM gateways. SIM Box fraud transports international calls through VoIP, then routes them to the local cellular network via a collection of SIM cards inside a SIM Box device.

To explain how SIM Box fraud is committed, firstly we describe the legitimate way for international calls. Let us assume that caller A and caller B are in different countries. Caller A makes a call to caller B over the mobile operator. The mobile operator of country A takes the call and sends it through its international gate to a transient operator. The transient operator then routes the call through voice over IP (VoIP) to country B's mobile operator and pays a toll. After that, the mobile operator of country B terminates the call through its network to caller B. Figure 1 [15] shows the legitimate route.

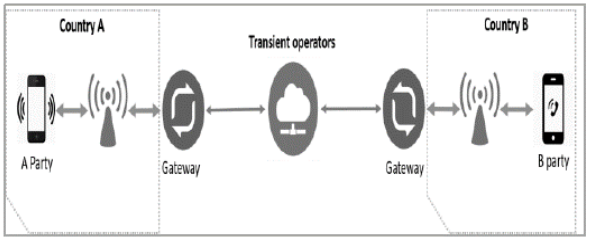


Figure 1: Legitimate Route of International Call

In bypass fraud, the transient operator routes the call through a SIM Box placed in country B using VoIP, the SIM Box then reroutes the call through country B's mobile operator and pays for just the

local call. Figure 2 [15] shows the bypass fraud route.

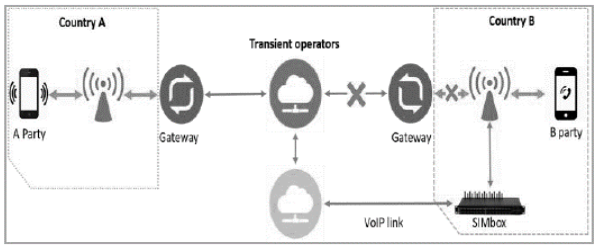


Figure 2: Bypass Fraud Route of International Call

The incentive here is that the toll charge by country B's mobile operator is much higher than the local call fee. Bypass fraud is committed when fraudsters install SIM Boxes with multiple low-cost, prepaid SIM cards. SIM Box equipment includes SIM slots and antennas. The SIM Box is connected to the Internet through an Ethernet port.

d. On-Net Vs Off-net Bypass Scenario

We can divide Bypass fraud into two major categories as On-net bypass and off net Bypass. On-net bypass means fraudsters use the SIM cards of same network of destination number. This is the common case in most of the countries as calls within same network cost much lower than charges for calls to other operators or international calls. Instead of using the costly high-quality routes that go through destination networks' International Switching Center (ISC), some wholesale carriers route calls through SIM Box operators connected through the Internet. The SIM Box dials that call via a SIM card of the same network of the destination number.

2.2 SIM Box Fraud Detection Techniques

Detection of SIM cards used for SIM Box fraud is a challenging task for mobile operators. The major methods used in battling SIM Box fraud currently in use include Test Call Generation (TCG), CDR analysis and controlling distribution of SIM Cards.

a. TCG (Test Call Generation)

TCG is used as an active method to detect bypass fraud where operators test different international routes to their network and analyze whether or not calls are coming through an international number or local number routes [16]. If it came from a local number, definitely it is associated with some SIM card used in a SIM Box that granted the fraud department to process it. TCG is one of the effective methods to detect SIMs used in SIM Boxes. However, this method costs the operator to generate test calls to the entire international routes. Also, since this method is based on generating calls to random numbers, getting SIM cards used by SIM Boxes is probabilistic.

b. Fraud Management System (Role -based analysis)

FMS is based on predefined rules and threshold that are manually defined by domain experts aimed to segregate SIM Box fraudsters. These rules and threshold are summarized features constructed from CDR data of subscribers. These are used to identify abnormal behaviour of SIM cards by comparing with analyzed subscriber role. Through this process it is able to distinguish fraudulent from legitimate usage. If abnormal subscriber behaviour is observed, then the system generates an alarm and experts perform analysis on it.

c. Non-technical Methods

SIM cards are indispensable to SIM Box fraudsters to survive since fraudsters need to maintain sufficient supply of SIM cards to be in business. However, SIM card distribution control will make this process difficult. Requiring government IDs and limiting the number of SIM cards per ID will prevent fraudsters from obtaining a large number of SIM cards to install in their SIM Boxes.

2.3 Challenges for Existing Defection Solutions

Fraudsters and anti-fraud are in eternal battle. Every time detection technology improves, fraudsters develop new methods to avoid detection and increase profit. This section describes different methods used by fraudsters to avoid SIM blocking.

a. Anti-Spam (Test Call Detection)

One of the effective methods to detect SIMs used in SIM Boxes is generating test calls (TCG) using different routes to known local network numbers. The incoming call will indicate weather it is coming from a local number or from an international number. If it was coming from a local number then it must be associated with some SIM card used in a SIM Box and easily processed by the fraud department. However, the fraudsters analyze the voice call traffic coming to their SIM Boxes and based on usage and other patterns they could determine whether the calls

were real subscriber calls, or they originated from a TCG system. They could then either block the test calls and prevent them from reaching the SIM Box or reroute the calls to a legitimate route to avoid detection.

b. Human Behaviour Simulation (HBS)

Smart SIM Boxes are designed to imitate the behaviour of normal customers by using HBS [14]. This technique makes detection of fraudsters very difficult if no advanced detection algorithms are used. HBS encompasses the following.

SIM Migration (Movability): Fraudsters deploy many gateways in different locations, for example, one in a city center and another in a shopping mall or some other crowded place and once in a while they swap the SIM cards between the gateways, so it would look like the user is moving. The swapping operation could be done manually or automatically using software.

SIM Rotation: SIM Boxes can be detected easily if fraudsters operate their SIMs around the hour excessively, so they limit their usage by rotation of the SIMs as workers shifts. This will make SIMs operate in limited hours a day, which simulates the behaviour of ordinary customers.

Usage of Other Network Services: Most of the SIM Boxes use just voice services and that makes them vulnerable to detection. In order to mitigate this issue smart SIM Boxes make calls and send SMS to each other. Also, sometimes they use some Internet services provided by the network operator. HBS makes dealing with SIM Box fraud harder and time consuming. Advanced measures must be taken to tackle this problem. In this work, real-time SIM Box fraud detection system is used to detect the bypass fraud by analysing CDR data.

3. Related Work

Various solutions were proposed by several authors to detect and prevent telecom fraud. A learning system has been one of the methods proposed for fraud detection tools [13]. These tools encode knowledge about fraud attacks as if-then rules which can represent and reason about some knowledge rich domain with a view to solving problems and giving advice.

The European Institute for Research and Strategic Studies in Telecommunications (EURESCOM) project, which was carried out in 2002, is the most significant contribution to the area of telecom fraud detection [13]. The project looked at how intelligent techniques from the domains of learning, personalization and data mining could improve fraud detection in a telecom service.

A technique that was proposed is the emerging pattern detection (EPD) [7]. EPD attempts to detect new forms of fraud by comparing two data sets, one with clean customer usage records and another one containing fraudulent records to identify patterns recurring in the second data set which were not present in the first data set.

A similar and powerful data analysis method is also proposed to discover SIM Box fraud patterns using integration of TCG and FMS [6, 7]. The suggestion is to incorporate TCG detection technology into the fraud management system. The integration attempts to detect new forms of SIM Box fraud using FMS CDR Analysis engine data sets by taking fraudulent behaviours of the already detected fraudulent records from TCG. In [7], a supervised learning algorithm is used to detect SIM Box fraud. The dataset is gathered from the operator and it includes CDRs from both legitimate subscribers and a fraudulent SIM Box. The proposed classifier has scored 98.7% accuracy in identifying the SIM cards that are used in the SIM Box device.

A similar study was done by Feven Fesseha [10], who used predictive model using data mining technology to detect SIM Box fraud. Twelve Selected CDR attributes were used to train the model. The study was conducted using WEKA software with five data mining algorithms for classification and the BFTree is chosen as the best to SIM Box fraud detection with an accuracy of 98%.

Frehiwot Mola [11] did a similar research on SIM Box fraud detection. The algorithms used to detect and predict SIM box fraudulent numbers are PART, J48, multilayer perceptron and hybrid. Using voting combination mechanism and by taking the strength of J48 decision tree and PART from rule based classification algorithms a hybrid algorithm is built. Nine features extracted from CDR data are utilized to

build a model that can be used to distinguish between legitimate and SIM box fraud calls. The feature includes calling number, called number, call fee, call duration, date and time, location number, total number of calls per day, SMS and GPRS call detail records. PART classification from rule based induction resulted in 99.49% accuracy and performs better than multilayer perceptron, J48 and Hybrid algorithms (J48 and PART).

In [12], the researcher analysed CDR data and identified twelve features to distinguish SIM-Box fraudulent from legitimate subscribers. The author proposed three dataset profiles (4 hour, daily, and monthly aggregated), to detect within possible short period of time. The researcher selected three classifiers: RF, ANN and SVM. The RF model scores slightly higher than the other two on the 4 hour dataset profiling. The accuracy of the model built with the 4 hour datasets are 95.98% and on the daily and monthly dataset is 98% and 99.05% respectively.

Tigistu Debebe [9] used ethio telecom as a case study. The objective is not detection but management principles on global fraud detection approaches. Finally the author proposed that ethio telecom needs advanced fraud management system with different technological options to detect fraud proactively.

4. The Proposed System and Findings

We have followed Lambda architectures to meet the research objectives. The Lambda Architecture is the new paradigm for big data that helps in data processing with a balance on throughput, latency and fault tolerance. The Lambda Architecture defines a set of analytic layers for building a complete big data system: Real-time Layer, Batch Layer and Serving Layer. Each layer satisfies a set of properties and builds upon the functionality provided by the layers under it. Figure 3 shows the high-level system architecture of the proposed real-time SIM Box fraud detection system.

The proposed detection system design was done in four stages:

a. Data Collection and Cleaning

Call Detailed Records were obtained from ethio telecom core systems. Basically there are two main data sources, namely, Local and International CDRs that are used in our design. Local CDR means transaction logs for calls originated by operator's own subscribers. These records are generated at Mobile Switching Centers (MSCs). International CDR records correspond to calls that originate to or terminated from foreign operators and detailed version of these CDRs are generated there. As seen from Figure 4, a new file is generated in less than a minute in MSC. Those CDRs files are originally generated as Abstract Syntax Notation One (ASN.1) formatted files and then converted to Comma-separated values (CSV) format for our processing. This data can then be blended with historical or other enterprise data sources including Equipment Identification Register (EIR) and history data from Fraud management database (FMS) to enrich the feed before further processing by downstream systems. The CDR's sensitive data like user numbers (MSISDN) or identities were modified. CDR data was processed in order to remove missing values, duplicate information and useless fields.

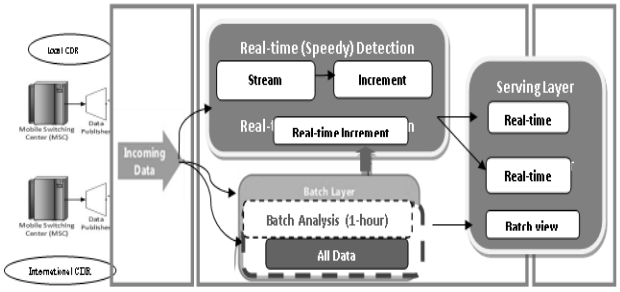


Figure 3: High-level System Architecture

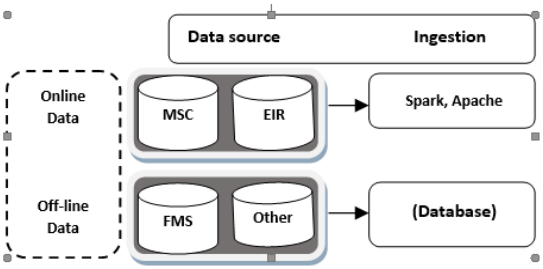


Figure 4: Data Sources and Ingestion

The real time Ingestion Service is the entry point to the streaming architecture as it decouples and manages the flow of information from MSC and EIR to the processing and storage components by providing reliable, high throughput, and low latency capabilities.

b. Feature Extraction and Engineering

To prepare the data for the machine learning model, informative features for each CDR were extracted. Also, features had been engineered in order to get the best features for training the model.

Every time a call is placed on a telecommunications network, descriptive information about the call is saved as CDR. CDR information can be used to analyse the characteristics of each call. The dataset used for this work are listed in Table 1.

Table 1: Data Set

	1-houd window	*Near Real-time (<2min)*
Training data	11573	11573 (375 separated by 2min)
Known fraud	333	141 in each 2min

First dataset comprises of 11573 CDR entries of which 332 were known fraud and the rest are for testing purpose. Further we have Split subsets of data to train the model, test it and further to validate against new dataset. All features of the CDR are not used directly for data mining, since the goal of data mining is to extract knowledge at the customer level. Thus, the CDR associated with customers must be summarized into a single record that describes the customer's calling behaviour. All relevant data for each subscriber was assembled in the form of columns and each raw dataset represents a unique subscriber. The choice of features is critical in order to obtain a useful description of the subscriber. So, a total of 12 features have been identified to be useful in detecting SIM Box fraud in 1-hour window and 6 future for near real time detection. Table 2 shows the list of these features and their corresponding data types. The selected features are based on the literature studied on the typical characteristics of SIM Box fraud subscribers as well as contribution from the experience of the staff who work on telecom fraud.

Table 2: Features Identified for 1-Hour Window Detection

Attribute	*Description*
Sub_No	Calling Number (Primary Key)
Idd_Call_Times	Toltal International Reciving Call
Idd_Calling_Times	Total International Outgoing Call Count Originated By Given Subscriber
Sms_Send_Cnt	Count Of Sms Sent By Number
Sms_Receive_Cnt	Count Of Sms Recived By Number
Idd_Sms_Cnt	Total International Sms Sent By Number
Calling_Duration	Total Outgoing Call Duration By Number
Cell_Count	Total Number Of Distinct Cell Coun By Number
Risk_Area_Coun	Number Of Orginating Number Risk Area Count
Bn_Missed_Idd_Calling	Count Of International Miised Call Made By Bn.
Calling_Imei _Cnt	Number Of Different Imei Numbers Used By Calling Party.
Is_Fraud	Known Fraud Numbers For Traning

We use custom ratio to split it into the two subsets, use the first 90% of data for training and the subsequent 10% of data for testing.

For the near real time detection, we rejected fields contain any instance of ***extreme usage*** related attributes. Extreme usage scenarios are relatively invalid compared to real-time SIM Box fraud instances. Therefore, we have taken CDRs for past instances of non-extreme usage related fraud cases and used those data to build our near real time logics.

Locating near real time patterns SIX CDR attribute are identified see table 3. Some attributes are selected from the two minutes collected CDR of unsuccessful call from the international CDR.

Table 3: Features Selected for Near Real Time Detection

Field Name	*Description*
Prod_ID	This is the Subscriber Identity Module (SIM) number which will be used as the identity field
MSISDN_Prefifix	number the First 5 prifix numbers
AN_Cell_ID_RISK_AREA	Cell site risk area flag
BN_IDD_SMS_in_2_min	The total international SMS initiated by BN with in 2_min
BN_Unsuccessful_IDD_Call	Total unsuccessful international Calls made in 2 minute
AN_Cloned-IMEI_Flag	AN_IMEI registered fraud flag
IS_Fraud	Known fraud numbers for training purpose

c. Model Training

The features extracted in the previous stage were used to train machine learning models. In this stage,

we used two different models. Besides real time analysis, we trained one-hour detection scenario to detect SIMs that failed to be detected in the real time scenario. Decision Tree, Extra Trees Classifier (random), Logistic Regression and Random Forest, Linear SVM, MLP Neural Network and Gradient Boosting Classifier had been used as supervised learning algorithms to train the model.

d. Model Evaluation and Testing

The final stage was testing and comparing the performance of the designed algorithms in terms of accuracy. To achieve this objective we used stream mining algorithms for near real time and IBM spss modular for one-hour decisions making. Figure 4 shows integrated stream mining to be applied for the proposed SIM Box fraud detection solution.

The training dataset used for real time SIM Box fraud detection consists of 11752 distinct CDR records and 331 known fraud cases. After several cycles of fine tuning, the model fits to detect. The test dataset contained all 5670 subscribers out of which 155 are known fraudulent calls. The proposed system detected 154 fraud instances correctly with one false positive and 12 false negatives. The same data sets with different features were used and out of it 139 frauds were detected using near real time detection scenario whereas the remaining 15 additional detections were made using one-hour window detection. As a result, near real time detection scenario acquired highest percentage of 81% of detection rate whereas one-hour window scores 99%. Detected in near real-time before huge impact is felt to operator with 98.5% over all detection accuracy. Following the research result, we consulted telecom fraud subject experts. They confirmed that the research result is promising to support telecom operators by detecting SIM Box fraudsters in a very short time.

- Linear SVM classification model is chosen as the best by IBM modular classifier. It scored 99.61% accuracy. This algorithm examined all the 11 fields in the training data set.
- VFDT classification model is chosen as the best real time classification by AutoAI IBM classifier. It scored 98.5% accuracy. This algorithm examined all the 6 fields in the training data set.

5. Conclusion and Future Work

The focus of this research is to detect SIM Box fraud in telecom networks in near real-time using streaming analysis techniques on call patterns indicated from the CDR. As a result, we followed the Lambda architecture and developed a system architecture that comprises batch (1-hour) and real-time layers, as it is well suited for applications which perform both near real-time and batch analytics. IBM cloud solution was used for both batch and real time SIM Box detection experiment due to its ability to perform high-speed processing. As a result, we have significantly increased the detection speed of SIM Box fraud. As per our observations, the proposed system performed the needed functionality.

As a future work, more refinement of data features will be done to improve performance and accuracy of the real time detection technique, for instance, including very sensitive data like signaling data will highly advance the detection capability.

References

[1] Global Fraud Loss Survey, C. F. C. Association et al., “Global fraud loss survey,” Press Release, http://www.cfca.org/fraudlosssurvey.

[2] A. Abdallah, M. A. Maarof, and A. Zainal, “Fraud detection system: A survey,” Journal of Network and Computer Applications, Vol. 68, pp. 90–113, 2016.

[3] M. Yelland, “Fraud in mobile networks,” Comput. Fraud Secur., Vol. 2013, No. 3, pp. 5–9, 2013.

[4] James Le, 2017, “Blue Ocean Thinker” https://jameskle.com/.

[5] P. Gosset and M. Hyland, “Classification, detection and prosecution of fraud in mobile networks,” Proceedings of ACTS mobile summit, Sorrento, Italy, 1999.

[6] Airn Vipin, ”Analysis and detection of SIM Box”, 2018. www.ijariit.com Combating Against Telecom Fraud: http://www.mmtelcom.com/webdex/fraud_prev.html.

[7] Latro services, 2016, "Beat Fraud before Fraud Happens", http://www.latroservices.com/

[8] Farvaresh, H. and Sepehri, M. M., "A data mining framework for detecting subscription fraud in telecommunication", Engineering Applications of Artificial Intelligence, 2011.

[9] Tigistu Debebe, “Detecting fraud in a mobile telephony network, ethio telecom as a case study", Unpublished Masters Thesis, HiLCoE, 2017.

[10] Feven Fesseha, “Predictive SIM Box Fraud Detection Model for ethio telecom”. Unpublished Masters Thesis, HiLCoE, 2017.

[11] Frehiwot Mola, “Analysis and Detection Mechanisms of SIM box Fraud in The Case of Ethio Telecom”, Unpublished Masters Thesis, Addis Ababa University, 2017.

[12] Kahsu Hagos, “SIM Box Fraud Detection Using Data Mining Techniques: The Case of ethio telecom”, 2018.

[13] Bradley Reaves, Ethan Shernan, and Adam Bates, "Blocking Cellular Interconnect Bypass Fraud at the Network Edge", 2015.

[14] Murynets, M. Zabarankin, R. P. Jover, and A. Panagia, “Analysis and detection of SIM box fraud in mobility networks,” Proc. IEEE INFOCOM, pp. 1519–1526, 2014.

HiLCoE

School of Computer Science and Technology

HiLCoE, one of the first higher education institutions in Ethiopia, was established in July 1997. It is a specialized Computer Science and Technology College that adheres to computing and quality of education. It is now widely recognized that HiLCoE prepares students to become industry and academic leaders who can apply technology and computer science principles across a wide variety of fields.

HiLCoE, since its inception has a vision of being Center of Excellence in ICT Education, Research and Development (ER&D). To realize this vision, in addition to its role in academia, it offers cutting-edge consultancy in the areas of Information Systems Security, Enterprise Architecture, Strategic and Solution Architecture, Service Engineering, Software Development, Project Management, Distributed Systems and Mobile Applications and many more.

Opportunities are tremendous with attractive reward for every effort you put to know and implement the evergreen field of Computer Science and Technology. Your dreams become realized by choice to join this pioneer and **Center of Excellence** in ICT ER&D in parallel with cutting-edge consultancy.

A commitment to excellence, coupled with continuous improvement, will result in HiLCoE being recognized as innovative, dynamic, and exciting community to learn, teach, and work with.

HiLCoE offers:

Three undergraduate programmes:

- **BSc Degree in Computer Science** and
- **BSc Degree in Information Systems** (in progress).
- **BSc Degree in Software Engineering** (in progress).

Three postgraduate programmes:

- **MSc Degree in Computer Science**,
- **MSc Degree in Software Engineering**, and
- **MSc Degree in Information System Security** (in progress).

HiLCoE Computer Systems Engineering, from its industry link, offers **specialized tailored training and consultancy** in:

- Enterprise Architecture,
- Strategic IT Architecture,
- Information Systems Security, Policy and Architecture,
- Project Management,
- Systems Development,
- Cloud Computing, and
- Mobile Application and Computer Network Design and Implementation.

HiLCoE paves a fabulous highway that makes you capable to be partner and contributor to the development arena.