

HiLCoE Journal of Computer Science and Technology

December 2024

Volume 11

ISSN 3008-0916



HiLCoE

School of Computer Science and Technology

Editor-in-Chief

Mesfin Belachew, HiLCoE School of
Computer Science and Technology
E-Mail: hjcst@hilcoe.net;
mesfinbelachew@hilcoe.net

Editorial Board

Mesfin Kifle, Department of Computer Science,
Addis Ababa University
E-mail: journal@hilcoe.net

Tibebe Beshah, School of Information
Science, Addis Ababa University
E-mail:
tibebe.beshah@gmail.com

Copyright©2024 HiLCoE, Addis Ababa. All rights reserved. This Journal, or any part thereof, may not be reproduced in any manner without the written permission of HiLCoE.

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

All articles published in the *Journal* represent the opinion of the authors and do not necessarily reflect the official policy of HiLCoE, the Editorial Board or the institution with which the author is affiliated, unless this is clearly specified.

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

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

Primary Objectives:

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

Contents

A Comparative Study on Airborne LiDAR Data Processing Software for a Smarter Road Construction in Case of Space Science and Geospatial Institute Tsegamlak Molla	1
Exploration of the Potentials of Blockchain in Helping Assure Public Trust in the Registration of Eligible Citizens and Transfer of Ownership of Public-Funded Houses in Ethiopia Fesseha Zemenu and Taye Abdulkadir	20
Image Based Similar Product Recommendation System for E-commerce Platform Girum woldkidan and Tibebe Beshah	30
Challenges and Readiness for Cybersecurity Threats at Selected Private and Governmental Banks Betelhem Kidane and Tibebe Beshah.....	45
Predicting Individuals' Vulnerability to Social Engineering Attacks Tesfaye Abaya and Tibebe Beshah.....	54
Recommended Computer Aided Design Software for EASE Engineering PLC Asmerom Estifanos	68
Comparative Analysis of Citrix and VMware Horizon Virtual Desktop Infrastructure (VDI) Solutions in Organizational Environment Dawit Tekle and Fanthaun Bogale	80
Practicing ESB Technology for Internet, Mobile Banking and Third Party Services Integration in case of Tsehay Bank Baheredin Surur.....	98
Distributed Database Selection for Vital Registration System of Addis Ababa Mesfin Debebe.....	110
Communication from the Editorial Board of HJCST The Editorial Board	121

A Comparative Study on Airborne LiDAR Data Processing Software for a Smarter Road Construction in Case of Space Science and Geospatial Institute

Tsegamlak Molla Marye
School of Computer Science HiLCoE, Ethiopia
tsegamlakm@gmail.com

Abstract

The application of IT in the Ethiopian Construction industry has been quite questionable in the past depriving the country from enjoying the benefits. IT practices such as using Light Detection and Ranging (LiDAR) in surveying for road design is relatively new to the country. The current geometric design manual being used by Ethiopian Road Administration (ERA) doesn't direct the use of LiDAR technology for surveying, yet that hasn't stopped ERA from buying its LiDAR surveyed data output from Space Science and Geospatial Institute (SSGI). In this paper, available LiDAR point cloud post processing software that could be used by SSGI are investigated. The software are compared with a thorough analysis of the technical, organizational and business aspects at hand in order to recommend the best for SSGI.

KEYWORDS: Light Detection And Ranging (LiDAR), Ground Control Point (GCP), Inertia Measurement Unit (IMU) & Global Positioning System (GPS)

1. Introduction

The construction industry is one of the major contributors to the Ethiopian Gross Domestic Product (GDP). According to [1], Construction was one of the sectors that accounted for most of the country's GDP growth despite the COVID-19 pandemic. Quick, sustainable and critical advancements such by highly involving IT to bring about technological progress to the industry is vital to Ethiopia's socio-economic growth.

The Ethiopian Construction industry is lacking from several fronts that are depriving its growth, one being in performance [2]. It's identified in [3] that lack of strong IT application is one of the factors hindering the capacity of acting bodies such as consultants and contractors, which has contributed as a factor to its poor performance. Thereby foreign construction firms are taking up the majority of big and complex contracts over domestic firms. For example, 55% of the road sectors big and complex project contracts were awarded to foreign firms as reported in [4].

The road construction in Ethiopia has gradually increased, reaching 144,024 km of all-weather roads at the end of 2019/20 fiscal year. This accomplished the countries 41% road development goal of that fiscal year. The country had spent 37.3 billion birr in the previous

fiscal year on road development alone. By the end of 2023, it is envisioned by Ethiopian Road Administration to have 10,000km of road built in addition to existing using 41 billion birr. Ethiopian Road Sector Development Programs (RSDP) has been strongly pushing the development of new roads for the past 15 years. This demands the usage of latest technologies in civil engineering and IT to secure the intended goals for Ethiopian roads. [5]

The adoption of airborne LiDAR technology also known as Airborne Laser Scanning (ALS) in construction has revolutionized the industry. Latest ALS technologies boast of scanning at pulse repetition frequency of 2MHz, obtaining 5cm vertical accuracy by 13cm horizontal accuracy, flight heights ranging from 300m up to 5500m and thus acquiring 6 million points per second [6]. This accumulates vast amount of raw data especially for large road projects that is crunched by some LiDAR processing software in order to output usable data formats for instance orthomosaic or Digital Terrain Model (DTM). SSGI has been dealing with providing such remotely sensed survey data deliverables to its customers such as Ethiopian Road Administration (ERA), which is intended for constructing roads.

Road construction in Ethiopia according to ERA's manual in [7] recommends the use of Photogrammetry as one of its site surveying techniques, but not LiDAR for it is yet to be updated as disclosed by ERA's Road Design Director in the interview recorded in annex 2. Photogrammetry is another type of 3D scanning technology that uses photographic images to extract geometric information from an object [8]. Photogrammetry entails limited accuracy especially in cases of dense vegetation and edge detection affecting output quality [9]. Acquiring Photogrammetric data is also affected by insufficient day light; where by acquisition of data for instance is hindered on a cloudy day or limited to day time [9]. Thus the manual supplement's Photogrammetry with field survey (total station) seeking more reliable results, which concurrently diminishes performance due to reliance on traditional practices resulting in higher monetary and time costs as road project size increases. These limitations among many others pushed SSGI to using LiDAR technology over photogrammetry to provide site investigation data deliverables to ERA, bypassing the obsolete at hand ERA's 2013 manual. This change was welcomed by ERA as the LiDAR deliverables were also compatible to ERA's CADD system; contingent on the used LiDAR data processing software's compatibility to a CADD system.

ALS data has increased radically in magnitude and complexity. Modern ALS complemented with Full-Waveform (FWF) data contains an additional up to 256 samples for each return pulse compared to the traditional ALS survey which collected four or more returns per transmitted pulse. With the restrained ability of computational processing power and hardware the above mentioned large and complex ALS data is a burden to compute. Therefore, the processing of ALS data takes up long periods of time especially with traditional sequence algorithms. [10]

This comparative study addresses SSGI's problem of owning expensive post point cloud processing software with processing speed issues and disk storage issues due to vast output as recorded in annex 1. Therefore, this study aims to unveil the best possible ALS data processing software, for SSGI, currently on the market allowing for robust, capable and efficient processing abilities without breaking the bank. It will exhibit the application of IT in the road construction industry with 3D mapping tools that are capable of making construction practices smarter; all in all seeking advancement with enhanced performance to the industry.

2. Objective

The main objective is to compare and contrast among airborne LiDAR data processing tools in order to identify which tool is overall a better choice with regards to the technical, organizational and business perspectives, and the specific objective is to recommend the LiDAR point cloud post processing software for SSGI that would aid it deliver an efficiently achieved, accurately processed and feasibly acquired output, all done without compatibility issues whether to SSGI nor to its client ERA; all in all suitably satisfying SSGI's LiDAR processing needs.

3. Related Work

The application of IT in construction has become increasingly common deploying new technologies such as Internet of things, big data, cloud computing, virtual reality (VR), augmented reality (AR), unmanned aerial vehicles (UAV), three-dimensional (3D) scanning, geographic information systems (GIS) and digital construction [11]. 3D scanning aided by software is a non-contact measuring technology through digitizing 3 dimensional scans of physical objects allowing for design, analysis and development as mentioned in [12]; in addition to visualization capabilities. It cuts down time and effort in doing conventional contact measurements inevitably enhancing performance [12]. Photogrammetry and laser scanning are example types of such 3D scanning of which Laser scanning or LiDAR takes precedence in precession in times of dense vegetation, poor lighting and atmospheric conditions [13], [14].

Light Detection and Ranging (LiDAR) as defined by [15] is a remote sensing technology that sends out pulses of laser light and measures the time the reflection from the target object is detected back to the sensor. Each beam of light reflected back to the sensor creates a three dimensional point and a collection of those elevation points creates a point cloud forming a 3D representation of the earth's surface and its features. One way to collect large LiDAR data is by attaching the Laser scanning instrument to an aircraft, which is called Airborne Laser Scanning (ALS) [16]. As the LiDAR data is collected on a moving platform, it consists of devices such as global positioning system (GPS) navigation devices and inertia measurement units (IMU) are used to create a georeferenced point

cloud; as illustrated in Figure 1 [17]. The LiDAR sensor collecting such data is usually mounted at the bottom of the aircraft. This technology allow surveying areas that are large, difficult to access, unsafe for people, tough terrain, etc. that would be time consuming and costly to do using the traditional methods such as total station. This rendering of 3D reconstruction and mapping the scene aids the engineers and management with more effective and efficient decision making and work.

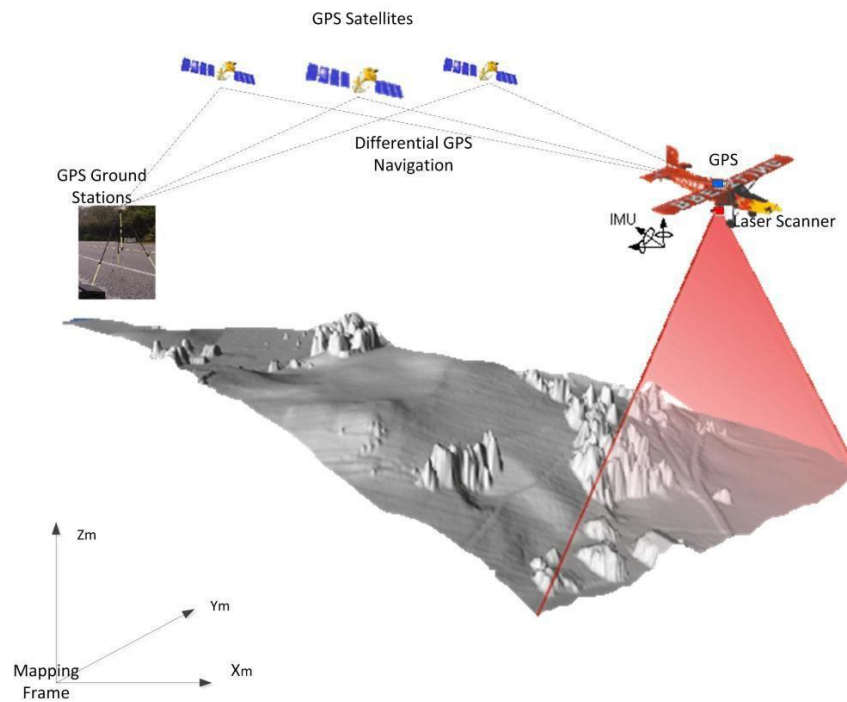


Figure 1 Airborne LiDAR data collection [25]

Field surveying and passive sensing approaches was the way to go for road geometry prior to the arrival of Laser scanning with all its bells and whistles. The traditional 2-dimensional terrain mapping was insufficient for road design, construction and management, which required site data to be in 3 dimensional formats. The conventional approach to acquiring 3D models done by surveyors were through manually digitizing aerial images with the use of few sparsely located ground control points, which took up a lot of time, was expensive and compromised the 3D model's accuracy. These limitations were overcome by ALS. [18]

Initial applications of airborne LiDAR were started by NASA in the 1970's working on aerial prototypes for "space borne sensor deployment". The remote sensing focused on environmental and terrain mapping studies. LiDAR has continually evolved into various applications rendering high accuracy as demonstrated by the German Stuttgart University

in geometrical profiling. Airborne LiDAR required a GPS system with full working area coverage and an accurate IMU system, which was made possible after mid-1980. The capacity of LiDAR to provide terrain mapping data had increased by mid-1990's providing from 2000 up to 25,000 pulses per second. Having gone through several steps of evolution LiDAR's high speed, great accuracy and active data extraction, while having market price stability, has made it an optimal solution for mapping. [19]

Advantages of ALS for road site surveying include but not limited to for instance, giving a wide range exposure on surveyed site sufficient for any geometric road analysis [18]. Second, ALS is not hindered by any obstacles on the site ground preventing it from have a clear view for scanning [18]. Third, the safety of both surveyors and drivers is not compromised as there is no interference with vehicles on the road nor is there any physical presence of surveyors on sloppy, tough and inaccessible terrains that might be a safety hazard [18]. In addition, it has the ability to map the ground beneath dense vegetation by penetrating laser pulses through minute gaps in the foliage [20]. Moreover, LiDAR is not affected by any external lighting outside of the sensor while collecting the data, for it's an "active sensing technology" [21]. This property ensures that scanning can take place at any time during day or night and under any "bright light or geometric distortions"; that would have commonly halted a Photogrammetric approach [21]. Hence, ALS is the optimal site investigation technology.

The main product found after processing airborne LiDAR data is a 3D model of earth elevations. This modeling of elevation allows for "site planning, site grading, earthwork, quantity estimates, and more", which means having a visual of the project site and ability to take action remotely without being physically present on the site [22]. The 3D earth models are classified as Digital Elevation Model (DEM), Digital Surface Model (DSM) or Digital Terrain Model (DTM). These models serve various purposes such as soil mapping, land slide studying, hydrological studying, vegetation management, aviation purposes, road development and management among others. A DSM captures all surfaces whether it's ground, vegetation or manmade surfaces above ground such as buildings or power lines [23]. DEM, on the other hand, models the pure ground points negating all other above- ground surfaces [23]. However, DEMs are further divided into more types dedicated to specific operations. For instance in the transportation community, the DEMs would include the bridge decks, though it is above-ground level, in order to have a complete view of the roadway system [24]. Lastly, DTM refers to 3 dimensional topographic model of the bare- earth consisting of a collection of points with reference to a geodetic datum, thus equivalent to adding vector features to a DEM like "ridges and breaklines" [24]. This model holds the vital bare-earth topographic representation used by the highway engineering sector [25].

Though ALS technology has shown significant growth in 3D data collection and

quality, it is yet to relieve its users working on point cloud processing from time intensive manual correction of classification errors and lack of full automatic feature extractions [18]. Classification of surface features fall into ground and above ground categories, whereby the ground surface is “the continuous and smooth surface that has nothing visible below it” whereas above ground objects are further classified as attached or detached objects. Attached objects elevate above bare-earth on certain sides but not all like bridges and viaducts, while detached objects are elevated above the bare-earth surface on all sides like buildings and trees. [25]

Filtering ground point from non-ground points is a necessary preprocess step required before applying interpolation on extracted ground points to generate 3D models such as DTMs [26], [27]. There are various filtering approaches or algorithms that yield different levels of efficiency depending on types of terrain. Hence a user must be versed with the various assumptions taken in the various ground filters in order to properly choose and apply them [26]. These algorithms are the engines running the numerous commercial software used to process our LiDAR data such as Axelsson’s algorithm that is implemented in TerraScan as reported in [27], which is the core module of the Terrasolid software suit that is currently being used by SSGI.

In the process of algorithms filtering ground points from non-ground points they consider preregistered suppositions of ground features. However, these assumptions are commonly mistaken by the actual ground characteristic in cases of shrubs, short walls along walkways, bridges, buildings with different size and shape, hill cut-off edges, complex mixed covering, areas combining low and high-relief terrains, and lack of reliable accuracy assessment. Small hedges that dot the ground surface when seen from above are called shrubs. They are usually less than a meter in height, but they are one of the most common ground features wrongly identified as ground. Short walls are another reason to mislead filters into assuming its ground especially when the LiDAR scanning is taken in urban places at high resolution. Likewise bridges also are reason for confusion to some filters such as single dimensional scanning filters, as bridges have certain sides directly elevated above ground while remaining sides linked to the ground surface in a continuously even manner. [26]

Buildings with different sizes and shapes may be wrongly labeled as ground in sections where it’s still part of the building. In addition to buildings other mixed coverings to terrains that happen mainly in urban areas persist to confuse filtering algorithms [26]. Hills with sharp cut-off edges also end up getting mislabeled as above ground in areas the slop exceeds the highest ground slop limit [26]. Moreover, filters run into the tough dilemma of choosing among slop or elevation limits in areas of varying relief between high and low [26]. Hence choosing commercial software to accurately process LiDAR data with filters that don’t fail to recognize the pitfalls of mislabeling terrain is a tricky task.

Road construction and maintenance is done all over Ethiopia in both rural and urban areas. As mentioned above urban areas are quite challenging to ground and non-ground filtering algorithms as they have highly mixed coverings over the terrain, which can be confusing to the filters. In addition to mixed land cover features over urban terrains, rivers that cross such urban areas also play a factor in filters miscalculating ground points during interpolation; creating inaccurate DTMs [27]. In the Ethiopian context, this will certainly pose a problem as rivers flow in all areas of the country, whether urban or rural.

Various landscapes require different filtering approaches suited to the respective landscape. Applying similar slope configuration for both areas that consist of hilly and flat landscapes may conclude in an inaccurate ground point classification [27]. This highlights the importance of filters to be adjustable allowing for parameters to be tuned to the terrains context for more accurate results [27]. Iterative filtering approaches are considered to yield tunable filters such as Axelsson's algorithm [27]. Filters with adaptive ability such as to slope thresholds is another approach proposed by [27].

Different types of filters that generate varying qualities of DTM for either specific or general terrains have been developed. "Block-minimum, slope operator, triangulated irregular network (TIN) modeling and raster calculation" are some approaches [28]. While another comprehensive review of filters compiled by [26] identifies 6 types or categories namely segmentation and cluster-based, morphological, directional scanning, contour-based, TIN- based and interpolation-based filters. The paper in [26] demonstrated and analyzed typical algorithms for each category with their respective advantage and disadvantages. It continued to examine the capabilities of 10 commercial ground filters in 3 various terrains of firstly a site with rough slope and dense vegetation, second with relatively flat urban areas containing objects of various sizes and shapes, and lastly a site with rough terrain and discontinuous surfaces. The conclusion of [26]'s study suggested that the ground filters were most challenged by surfaces with rough terrain or discontinuous slope, dense forest canopies and regions with low vegetation.

Another most recent review on ground filtering algorithms in Z. Chen et al. [28], which having considered existing filter algorithm approaches, managed to categorize them into 6 types consisting of surface-based adjustment, morphology-based filtering, TIN-based refinement, segmentation and classification, statistical analysis and multi-scale comparison. The paper was able to compile the best terrain suitable and not so suitable for each filter type along with the memory storage demand and computational efficiency of each of the 6 categories.

Thus, the summary of the 6 categories made in [28] is as follows starting with the first, surface-based filtering method is suitable in forested areas while should be avoided in areas of rough and steep terrains. It has a high demand on memory storage but a middle

level computational efficiency. Secondly, the morphological based type can be arguably used in steep terrains and terrains with very little mixture of object coverings while not being suitable in complex mixed covering terrain. It requires very little memory storage while yet having a high computational efficiency. Thirdly, TIN based is said to be suitable for steep terrains, yet unsuitable for urban areas and discontinuous terrains, while having a middle level demand on memory storage and also middle level computational efficiency. The fourth segmentation-based filter type is meant to be used in urban areas and terrains with various objects, but not in areas of rough and steep terrains or dense forests. However, the paper [28] didn't have data on memory storage demand nor computational efficiency for segmentation-based filters. Fifth, statistical analysis is recommended to be used in generally flat terrain but not in terrains with various objects, and it's a low-level demander in memory storage with also a low-level computational efficiency. Lastly, multi-scale comparison is best suited in urban areas while not recommended in rough and steep terrains, with a memory storage demand of middle level and a low computational efficiency.

The above study shows that each filtering method specializes in certain type of areas but not in others. The filters best suited in urban areas might not work for rural forested areas [28]. In [28], the paper suggests combining the filters to get the benefits of multiple filters. It also suggests combining the generated DTMs from various filters after the filtering process has taken place, also to extract the benefits of various DTM generators post the filtering; thereby increasing accuracy [28].

The entire filtering algorithms mentioned above fall into the rule-based filtering method category. Rule based algorithms put into consideration the physical elements of point clouds in order to distinguish non-ground from ground features; normally done with a workflow of procedures [29], [30]. Despite not having a one-size-fits-all rule-based algorithm there has been progress in development of machine learning algorithms that are generally applicable to all terrains. There are two classes of machine learning algorithms, supervised and unsupervised machine learning algorithms [30]. Supervised machine learning is further divided into two, whereby the labeling data classes are done manually or automatically with the use of automatically generated training data [30]. On the other hand, training data is not needed for unsupervised machine learning, which is further subdivided into 4 [30]. They are classification tree method, grouping and separability methods, and rule application methods [30].

Whether the underlying development to filtering approach of the commercial software is rule based or machine learning one thing is clear, the LiDAR processing software performance remains to be a point of research. [31]

4. Comparative Analysis and Detailed Recommendation

4.1 Comparative Analysis

This comparative study has taken the technical, organizational and business aspects into consideration in determining the better software. This research has applied dynamic software testing mechanisms for the technical comparison aspect to choose the best Airborne LiDAR data processing software for SSGI. Dynamic software testing entails running the program under scrutiny unlike static testing techniques which don't [32]. As we expand Dynamic testing category we find Functional and Non-functional testing techniques, whereby Functional testing types are further subdivided into Black-box, White-box, Grey-box, Unit testing, Integration testing and more [32], [33]. On the other hand Non-functional testing techniques refer to Performance, Security, Usability and more testing methods [33]. This study has considered one of each, meaning Black-box testing from Functional and Performance testing from Non-functional software testing techniques.

As for Black-box testing, this study has chosen to apply a subtype technique identified in [34] as Use Case testing technique to determine the functionality in the short-listed software. In addition, the level of accuracy particularly vertical accuracy sustained by the GIS software while generating digital terrain models by interpolating the input point cloud has been measured by Root Mean Square Error (RMSE). RMSE is one of the statistical considerations under taken to determine vertical accuracy of LiDAR data [35]. RMSE can be defined as follows:

$$RMSE = \sqrt{\frac{\sum_{i=1}^n (Z_{data(i)} - Z_{check(i)})^2}{n}} \quad [35]$$

where $Z_{data(i)}$ refers to the actual elevation data at check point (i), $Z_{check(i)}$ refers to the elevation data determined by the GIS software after inputting the LiDAR point cloud at check point (i), and lastly n refers to the entire collection of check points or also known as ground control points (GCP) [35]. This study has taken the max limit of 12.5 cm RMSE as the standard limit for acceptable accuracy according to [36], which refers to the Idaho LiDAR Consortium's (ILC) standards. Therefore, the software with the lowest RMSE that is also under the 12.5cm limit would be distinguished as the most accurate.

Regarding Performance testing, measuring points of response time (processing time taken), resource utilization such as Central Processing Unit (CPU), as well as exerted strain on Random-Access Memory (RAM) and disk was measured and compared among the software, which are identified by [37] as key performance indicators (KPI) under efficiency-oriented indicators and service oriented indicators respectively of a performance test. The performance test consists of testing for achieving basic GIS processing tasks on

LiDAR data such as loading point cloud data, making classification particularly ground classification, which also includes filtering the point cloud data accordingly, applied in all candidate software. This test will be carried out with the use of Windows Performance Recorder (WPR) and Windows Performance Analyzer (WPA) along with the respective candidate software.

The second main aspect of this comparative research is the organizational aspect. It took into consideration SSGI's computing capacity and looked into its ability to handle the short listed software. It also compared the softwares' ability to interoperate with CAD systems and import LAS formatted point cloud. Lastly, the business aspect of this comparison was further investigated to acquire the most feasible or cheapest software, yet one with the best value for money invested without breaking the bank.

Initially the list of software to compare was 11, with software namely Terrasolid, Global Mapper, TopoDot, Trimble, VRMesh, LiDAR360, Phoenix, Mapteks PointStudio, Arc Gis Pro, Vision LiDAR, and MATLAB. Firstly, Terrasolid was chosen as part of this study for the obvious reason of its current use by SSGI. Secondly, Global Mapper was selected for the reason of it being a familiar tool to the SSGI staff, though it's only being used for visualization purposes with its unregistered post free trial version, with all advanced features turned off, there could be technical capabilities not explored in its full paid version. Trimble Realworks and TopoDot were also selected for they are the most popular softwares comparable to Terrasolid as seen in [38]. The remaining softwares were selected for reasons of being high tier LiDAR post processing softwares. However, the trial version for demonstration purposes was acquired for only 4 of the listed softwares. The demo software was acquired for namely Terrasolid, LiDAR360, Mapteks PointStudio and ArcGis Pro. The comparison is thus focused on comparing the 4 LiDAR point cloud post processing software.

4.1.1 Detailed Technical Aspect Comparison

The technical comparison comprised of applying a Black box Use Case testing approach for Functional testing method and Performance testing approach for Non-functional testing technique as mentioned above. The software tests and comparisons were accomplished by using the point cloud and GCP data provided by SSGI on all short listed GIS software. The point cloud data was scanned from the Entoto to Kotebe area in Addis Ababa and consisted of 51,753,630 points covering a mixed type of terrain. A single laptop was used, having an Intel(R) Core(TM) i7-8550U CPU at 1.80GHz & 2.00 GHz processor, 24GB RAM and a 64bit system platform that is running a Microsoft Windows 10 Pro operating system.

Starting with the Black Box Use Case testing, there are 2 Use Case models used to make the comparison. There are 7 set of basic necessary steps performed as mentioned by

SSGI's GIS technician interview shown in annex 1. The difference between the 2 Use Case models as seen in table 1 below happens at the 4th step, where GCP's that are out of the perimeter of the input point cloud or scanned area are included in use case 2 to possibly mislead the software.

Step	Use Case 1 Tasks	Use Case 2 Tasks
1	Import LAS format raw point cloud data acquired form SSGI	Import LAS format raw point cloud data acquired form SSGI
2	Perform Ground classification with filtering of the point cloud	Perform Ground classification with filtering of the point cloud
3	Generate a Digital Terrain Model (DTM)	Generate a Digital Terrain Model (DTM)
4	Import actual Ground Control Points (GCP) within scanned area	Import Ground Control Points (GCP) with additional off-grid check points
5	Calculate RMSE	Calculate RMSE
6	Make elevation adjustment	Make elevation adjustment
7	Recalculate RMSE	Recalculate RMSE

Table 1 Testing Use Case Modules

The resulting Black Box test shows Arc Gis Pro failed the test after the 4th step in both use cases, whereas the remaining softwares passed all steps in both use cases which could be seen in table 2 below.

Software	Pass/ Fail						
	Step 1	Step 2	Step 3	Step 4	Step 5	Step 6	Step 7
Terrasolid	Pass	Pass	Pass	Pass	Pass	Pass	Pass
LiDAR360	Pass	Pass	Pass	Pass	Pass	Pass	Pass
Mapteks PointStudio	Pass	Pass	Pass	Pass	Pass	Pass	Pass
ARC Gis Pro	Pass	Pass	Pass	Pass	Fail	Fail	Fail

Table 2 Black Box Test Result for both Use Cases

Having done the Black Box testing, where by checking if the software can perform required tasks step by step, this study also measured the degree of accuracy at which it produced the output. Thus the resulting accuracy assessment of the software DTM output is shown in table 3. It depicts a large margin of difference between Mapteks PointStudio

and remaining Terrasolid and LiDAR360 software. The result also shows ARC Gis Pro could not generate an RMSE value for both use cases as it had failed to carry out steps 5 through 7 from both use cases. Additionally, Maptek PointStudio also failed the RMSE test as its value was above the maximum 12.5 cm.

Software	Use Case 1 Results		Use Case 2 Results	
	RMSE Before Adjustment (cm)	RMSE After Adjustment (cm)	RMSE Before Adjustment (cm)	RMSE After Adjustment (cm)
Terrasolid	6.7794	0.0460	6.7794	0.0460
LiDAR360	6.6970	0.0720	6.6970	0.0720
Maptek PointStudio	970.713	943.6370	3,678.5570	1,746.7110
ARC Gis Pro	Unknown	Unknown	Unknown	Unknown

Table 3 Result of Black Box Test

Secondly, the Performance test was carried out simultaneously with the Black Box test during steps 1 and 2, meaning when loading the point cloud data and performing Ground classification with filtering of the point cloud. The two tasks were selected as they are the most prominent moments when performance is lowest according to the interview in annex 1. It was carried out in the background to see how the software performed with regards to speed and resource consumption. The result of the performance is seen in table 4 and 5 below.

Software	Trial No.	Point Cloud Loading Time (min)	Program CPU Usage (%)	Total RAM Usage (MB)	Total Disk Usage (MB)
Terrasolid	1	0.550	9.090	13,687.441	906.640
	2	0.167	6.620	14,141.074	30.726
	3	0.183	9.690	14,177.508	35.468
Mean		0.3000	8.4667	14,002.0077	324.2780
LiDAR360	1	0.883	6.540	20,232.242	4,543.963
	2	0.867	6.550	14,256.430	4,398.158
	3	0.517	3.800	14,465.059	3,172.219
Mean		0.7557	5.6300	16,317.9103	4,038.1133
Maptek PointStudio	1	1.200	0.880	16,991.629	1,977.735
	2	1.867	0.350	21,810.176	3,737.484
	3	1.017	1.210	23,953.176	2,064.638
Mean		1.3613	0.8133	20,918.3270	2,593.2857
	1	0.983	5.910	12,795.285	139.475

ARC Gis Pro	2	0.383	8.230	12,612.602	52.830
	3	0.717	6.300	12,437.477	62.418
Mean		0.6943	6.8133	12,615.1213	84.9077

Table 4 Performance Test during Loading Point Cloud Data
(Step 1)

Software	Trial No.	Ground Classification & Filtering Time (min)	Program CPU Usage (%)	Total RAM Usage (MB)	Total Disk Usage (MB)
Terrasolid	1	6.117	34.350	19,645.445	68.271
	2	5.900	31.350	19,547.488	80.216
	3	5.983	32.430	19,696.844	111.395
Mean		6.0000	32.7100	19,629.9257	86.6273
LiDAR360	1	14.333	12.540	10,326.125	1,063.355
	2	23.267	13.180	10,178.578	170.699
	3	20.233	12.930	10,626.957	129.366
Mean		19.2777	12.8833	10,377.2200	454.4733
Maptek PointStudio	1	1.533	0.280	28,826.145	1,367.915
	2	2.550	0.540	30,834.355	100.291
	3	1.150	0.230	32,060.273	1,875.724
Mean		1.7443	0.3500	30,573.5910	1,114.6433
ARC Gis Pro	1	1.150	9.740	13,774.965	2,544.226
	2	0.817	9.820	11,988.793	1,841.791
	3	0.817	12.48	11,403.016	1,623.589
Mean		0.9280	10.6800	12,388.9247	2,003.2020

Table 5 Performance Test during Ground Classification & Filtering (Step 2)

When considering both ARC Gis Pro and Maptek PointStudio have failed in prior tests, disqualifying them from further comparison, the research could focus on Terrasolid and LiDAR360. Therefore, the performance test shows Terrasolid was the fastest and least disk consuming software compared to LiDAR360 in both step 1 and 2. LiDAR360 had clearly a better CPU usage in both step's 1 and 2, but RAM usage was a close call among the two.

4.1.2 Detailed Organizational Aspect Comparison

The main intention in looking at the organizational perspective is to see if the software integrates well with the designated company's working environment. In order for a LiDAR post processing software to be acceptable for SSGI it should first be able to operate under SSGI's computing capacity and secondly generate an outcome that can be compatible with its clients processing application. Therefore, firstly the software requirements of each prospective software as registered in annex 4 was compared with SSGI's computing

specification which is also registered in annex's 1 and 3. The resulting deduction after the comparison is seen in table 6.

Secondly, the ability of each prospective softwares output to be compatible with CADD systems was compared also as seen in table 6. This entailed looking into each software capability to exporting in DWG, DXF or XYZ formats as determined from annex 1. This research also has taken the liberty on making sure they are compatible with LAS point cloud data as it was the type of data received from SSGI as input for processing within this comparative research.

Software	Compatible to SSGI'S Computing Spec.	LAS Compatible (Input Data)	CAD Compatible (Output Data)
Terrasolid	✓	✓	✓
LiDAR360	✓	✓	✓
Maptek PointStudio	✓	✓	✓
ARC Gis Pro	✓	✓	✓

Table 6 Prospective Software Input and Output Compatibility

Therefore, all the short listed software are compatible to SSGI's computing environment, are able to import LAS data types and interoperate with SSGI's client CADD system. Thus all satisfying the organizational needs of SSGI.

4.1.3 Detailed Business Aspect Comparison

Having looked at the technical and organizations aspects for comparison this study also took care of looking into the prospective software's through the financial lens. Each of the 4 prospective software companies was requested to provide an annual subscription quote for 2023. The acquired price quotes for annual subscription by each prospective software company is listed in table 7 as seen below. The comparison showed LiDAR360 to be the cheapest of them all.

Software	Annual Subscription
Terrasolid	\$1,767.70
LiDAR360	\$1,500
Maptek PointStudio	\$12,420
ARC Gis Pro	\$6,250

Table 7 Price Quotation as of 2023

4.2 Recommendation

Having considered various determinants to compare the short listed softwares for SSGI, this paper has selected the most important of them all for SSGI. They are time, CPU, RAM and disk utilization, accuracy, software to system compatibility, software output to SSGI's client CADD system compatibility and annual subscription price point. ARC Gis Pro and Maptek PointStudio were eliminated from the comparison early in the technical assessment. This study has shown Terrasolid remains to be the better software on the market over the remaining LiDAR360 software. This was so because Terrasolid was able to provide more accuracy on its DTM output in addition to also saving processing time on average by 60% in point cloud loading and 69% during ground classification with filtering, which is critical to SSGI, also overall better in performance when compared to LiDAR360. Terrasolid had no compatibility issues, but as did all. LiDAR360 may be the cheapest software, but when compared to Terrasolid there is only a \$267 markup on annual subscription price, which is insignificant to SSGI as it's a government institution. Therefore, seeking better post point cloud processing software is not the best strategy going forward for SSGI at this current point in time. I hence recommend seeking other solutions that are inexpensive to solve the remaining issues of processing time and storage issues. These are resource management and hardware upgrading problems. Therefore, I recommend resource management, especially with regards to CPU, memory and the disk drives in order to get better performance. The current software's processes should be given priority when running over other non-system processes. Minor steps such as discarding obsolete files or moving them to external storage could free up space on disk helping the storage issues. Another simple disk management could be to extend the hard drive partition into unallocated spaces resulting in enlarging disk space. Moreover, upgrading SSGI's 32GB RAM to 64GB, also upgrading processing capacity is an inexpensive way of gaining more performance, solving SSGI's computing issues.

5. Conclusion

There is potential for the Ethiopian construction industry to grow into the desired high performing industry. This is being made visible as government institutions such as SSGI and ERA allowing the use of high end softwares, such as Terrasolid, to digitize the surveying side of road construction; applying IT. This study has made a thorough comparison with regards to the technical, organizational and business aspects in determining Terrasolid remains to be the better software from the short listed 4 softwares. Recommendations that are cheap and inexpensive are made to enhance the current software abilities, thus completely answering all remaining issues. However, there were limitations in this study of not freely being capable to acquire popular softwares like Trimble Realworks and TopoDot and also using a single sample input point cloud though having a sufficiently mixed terrain type it lacked the presence of a river, which could have further shown the softwares' ground filtering capability.

References

- [1] World Bank, "The World Bank in Ethiopia – Overview;" WorldBank.Org, <https://www.worldbank.org/en/country/ethiopia/overview#1> (accessed Dec. 17, 2021).
- [2] P. Almeida et al., "Shaping the Future of Construction - A Breakthrough in Mindset and Technology," World Economic Forum, Cologne, Switzerland, 2016.
- [3] D. Mengistu and G. Mahesh, "Challenges in developing the Ethiopian construction industry," African Journal of Science, Technology, Innovation and Development, 2019, DOI: 10.1080/20421338.2019.1654252.
- [4] ERA, "Road Sector Development Program -19 Years Performance Assessment Report," Ethiopian Roads Authority, 2016.
- [5] T. Tefera and Dr. F. Gebrehiwot, "Ethiopia - Country Commercial Guide: Roads, Railways and Logistics", International Trade Administration, <https://www.trade.gov/country-commercial-guides/ethiopia-roads-railways-and-logistics#:~:text=The%20Ethiopian%20transport%20sector%20is,heavily%20monopolized%20by%20the%20government> (accessed Jul. 28, 2022).
- [6] Leica Geosystems, "Leica TerrainMapper-2 flexible linear-mode LiDAR sensor," <https://leica-geosystems.com/products/airborne-systems/topographic-lidar-sensors/leica-terrainmapper-2> (accessed Aug. 4, 2022).
- [7] Geometric design manual – 2013, Ethiopian Road Authority Standard, 2013.
- [8] G. Konecny, Geoinformation: Remote Sensing, Photogrammetry and Geographic

- Information Systems, 2nd Ed., CRC Press; 2014, pp. 9-10.
- [9] Scout Aerial Group, "Strengths and limitations of Photogrammetry," Scout Aerial Australia, <https://www.scoutaerial.com.au/article-photogrammetry/#> (accessed Aug. 19, 2022).
- [10] V. Cao et al., "Toward A New Approach for Massive LiDAR Data Processing," 2015, DOI:10.1109/icsdm.2015.7298040.
- [11] Z. Yu et al., "Smarter construction site management using the latest information technology," 2018, DOI: 10.1680/jcien.18.00030.
- [12] M. Javaid et al., "Industrial perspectives of 3D scanning: Features, roles and it's analytical applications" 2021, DOI:10.1016/j.sintl.2021.100114
- [13] A. Gil et al., "A comparison between LiDAR and photogrammetry digital terrain models in a forest area on Tenerife Island," Canadian Journal of Remote Sensing, 2013, 39:5, 396-409, DOI: 10.5589/m13-047.
- [14] W. Muecke et al., "Terrain echo probability assignment based on full-waveform airborne laser scanning observables," 2010.
- [15] J. Carter et al., "Lidar 101: An Introduction to Lidar Technology, Data, and Applications," Revised, NOAA Coastal Services Center, Charleston, SC, USA, 2012.
- [16] Y. Hu, "Automated Extraction of Digital Terrain Models, Roads and Buildings Using Airborne LiDAR Data," Ph.D. dissertation, Geomatics Eng. Dept. Calgary Univ., Calgary, Alberta, 2003.
- [17] A. Vo et al., "Airborne laser scanning data storage and indexing: state-of-the-art review," International Journal of Remote Sensing, 2016, 37:24, 6187-6204, DOI: 10.1080/01431161.2016.1256511.
- [18] Y. Zhou et al., "Highway alignments extraction and 3D modeling from airborne laser scanning point clouds," in International Journal of Applied Earth Observations and Geoinformation, Elsevier, China, 2021, DOI:10.1016/j.jag.2021.102429.
- [19] BCC Research Editorial, "Brief History of LiDAR, Its Evolution and Market Definition," retrieved from <https://blog.bccresearch.com/brief-history-of-lidar-evolution-and-market-definition> (accessed Dec. 6, 2022).
- [20] N. David et al., "PATHWAY DETECTION AND GEOMETRICAL DESCRIPTION FROM ALS DATA IN FORESTED MOUNTANEOUS AREA," September 2009.
- [21] S. Gargoum and K. Basyouny, "A literature synthesis of LiDAR applications in transportation: feature extraction and geometric assessments of highways," in

- GIScience & Remote Sensing, Taylor & Francis Group, vol. 56, no. 6, 2019, pp. 864-893, DOI: 10.1080/15481603.2019.1581475.
- [22] H. Grace, "Elevation Modeling - the differences between DTM, DSM & DEM," retrieved from <https://support.plexearth.com/hc/en-us/articles/4642425453201-Elevation-Modeling-the-differences-between-DTM-DSM-DEM>, (accessed Nov. 27, 2022).
- [23] GISGeography, "DEM, DSM & DTM Differences – A Look at Elevation Models in GIS," retrieved from <https://gisgeography.com/dem-dsm-dtm-differences/> (accessed May 29, 2022).
- [24] H. Heidemann, "Lidar base specification," in U.S. Geological Survey Standards, ver. 1.3, USA, 2018, ch. B4, DOI: 10.3133/tm11B4.
- [25] Y. HU, "Automated Extraction of Digital Terrain Models, Roads and Buildings Using Airborne Lidar Data," 2003.
- [26] X. Meng et al., "Ground Filtering Algorithms for Airborne LiDAR Data: A Review of Critical Issues," in Remote Sensing, 2010, vol. 2, pp. 833–860. DOI:10.3390/.
- [27] J. Susaki, "Adaptive Slope Filtering of Airborne LiDAR Data in Urban Areas for Digital Terrain Model (DTM) Generation," in Remote Sensing, 2012, vol. 4, pp. 1804-1819, DOI:10.3390/rs4061804.
- [28] Z. Chen et al., "State-of-the-Art: DTM Generation Using Airborne LIDAR Data," in Sensors, 2017, vol. 17, no. 12, DOI:10.3390/s17010150.
- [29] S. Ayazi and M. Seresht, "Comparison of traditional and machine learning base methods for ground point cloud labeling," in The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences, 2019, vol. 42, pp. 141-145.
- [30] Z. Gharineiat et al., "Review of Automatic Processing of Topography and Surface Feature Identification LiDAR Data Using Machine Learning Techniques," in Remote Sensing, 2022, vol. 14, no. 19.
- [31] V. Badenko et al., "COMPARISON OF SOFTWARE FOR AIRBORNE LASER SCANNING DATA PROCESSING IN SMART CITY APPLICATIONS," 2019.
- [32] T. Hamilton, "Static Vs Dynamic Testing: Difference Between Them," retrieved from <https://www.guru99.com/static-dynamic-testing.html#:~:text=Static%20testing%20is%20about%20the,testing%20is%20performed%20after%20compilation>, (accessed Feb. 16, 2023).

- [33] S. Kumar and A. Singh, "Software Testing," India: Vandana, 2019.
- [34] A. Roman, "A Study Guide to the ISTQB® Foundation Level 2018 Syllabus," Switzerland: Springer Nature Switzerland AG, 2018.
- [35] X. Liu, "Accuracy Assessment of Lidar Elevation Data Using Survey Marks," 2011.
- [36] M. Khanal et al., "Accuracy Comparison of Aerial Lidar, Mobile-Terrestrial, and UAV Photogrammetric Capture Data Elevations Over Different Terrain Types," 2020, DOI:10.3390/infrastructures5080065.
- [37] I. Molyneaux, "The Art of Application Performance Testing," 2nd Ed, CA: O'Reilly Media, 2014.
- [38] Y. Kin, "Automated LiDAR Extraction Software," in Caltrans Division of Research, Innovation and System Information, 2021.

Exploration of the Potentials of Blockchain in Helping Assure Public Trust in the Registration of Eligible Citizens and Transfer of Ownership of Public-Funded Houses in Ethiopia

Fesseha Zemenu
fzemenu@gmail.com

Taye Abdulkadir
HiLCoE School of Computer Science and Technology
tabdulkadir@ieee.org

Abstract

As a concept, “trust” is abstract and generally subjective. When there is mistrust between two entities, a third party trusted by both entities is usually involved to build the bridge of trust. This paper [8] sets out to explore the possibility of adapting blockchain technology to serve as the “trusted third party” between untrusting entities. As a case for study, there exists an issue of mistrust between the Ethiopian government (specifically Addis Ababa City Government) and the constituent public in the transparency of the process of registration of eligible city residents and transfer of ownership of public-funded condominium houses for registered, lucky lottery winners in successive rounds. In order to address this problem, the paper expounds the concept of trust first and goes on to explore blockchain technology in relation to its fundamental work in bitcoin. It compares four blockchain platforms, namely, Bitcoin, Ethereum, Hyperledger Fabric, and Multichain, and selects one for prototype implementation based on pre-defined evaluating metrics. The metrics used for the comparison are the platform’s customizability, maturity, performance / scalability, security, consensus algorithm, and relevance to the problem under study. The selected blockchain platform is then used for a prototype Proof of Concept (PoC) demo implementation to serve as the main element for assuring public trust in the transparency of the process for registration of residents and transfer of ownership for Addis Ababa Housing Project. Based on the results of the exploratory and comparative analysis, Multichain blockchain platform is selected and the Trust Assurance application prototype system is implemented in the cloud on AWS (Amazon Web Services). The Multichain-based PoC demo showed that the Housing Project data records can be made available publicly, that the system can’t be taken offline without turning off all nodes in the distributed peer-to-peer blockchain computing network, that the project data can’t be deleted or modified at will, that timestamps are an inherent part of every record, including who published them, and that the chronology of records can’t be altered. With the full consent of a Citizen’s Board and the Government’s goodwill in cooperating with the Board, the results of the PoC demo indicate that the public could trust the Housing Project data records if an enterprise-grade blockchain system is implemented in accordance with the procedures set out in this paper.

Keywords: Trust, Blockchain, Multichain, Addis Ababa Housing Project

1. Introduction

There are certain abstract human issues that technology may not directly address and solve, such as the issue of trust (or rather the lack of trust) that ordinary citizens may have on the activities of their government. A research article [1] from Deloitte outlines that a citizen's trust in government can be defined as citizens' confidence in the actions of a government to do what is right and perceived fair. The research suggests that governments can build and sustain citizen trust by focusing on four areas: humanity, transparency, capability, and reliability.[1] A key finding from Deloitte's research is the capacity of technology and innovation to drive change and to play a key role in trust transformations. Taking transparency as one of the four pillars in building trust between governments and their constituencies, ICT can play a significant role in bridging the trust gap.

One of the areas citizens have big questions of trust in the case of Ethiopia is in the conduct of government-led housing projects. The issue of transparency in this area has been a concern for a long time. Advances in information technology, particularly in innovative applications of blockchain concepts, have shown that technology can help assure citizens on the factual accuracy of verifiable government activities and records that cannot be easily tampered with. Government-led housing projects, in the case of Ethiopia and particularly that of Addis Ababa, are usually surrounded by suspicions and controversies. The public in general has a number of doubts and complaints about the conduct of the projects from registration of eligible citizens up to ownership transfer of housing units to citizens. Although the allocation of houses is done through a publicized lottery scheme, people still have doubts whether all houses are indeed transferred only to those who have been properly registered and only to those who have been lucky winners in successive rounds of lottery draws. It is this issue of lack of trust that this paper tries to address and suggest a mutually acceptable technological solution for the recording of the registration of eligible citizens for condominium houses and the subsequent ownership transfer of houses, that's secure, unchangeable, undeletable, incorruptible, and accessible by all parties for unequivocal verification of the whole process.

The following research questions (RQ) are formulated to address the above stated problems.

- RQ1: Can blockchain technology be used to record basic data about users registered for condominium houses, data about the housing units, and data about the results of successive rounds of lottery draws? If so, can the data thus recorded be kept tamper-free, unchangeable, and stored for perpetuity as promised by blockchain technology?
- RQ2: Is it possible to provide the immutable recorded information related to registered users, condominium houses, and corresponding lottery results accessible for public scrutiny?
- RQ3: How could blockchain technology help in earning the trust of the public and that the public gets assured "beyond any reasonable doubt" that the government-provided record of the housing projects is genuine, and that it couldn't be modified by anyone, even by technical and non-technical people in government authority, and that the information remains available and visible for anyone at any time?

2. Related Works

A research paper entitled "Secure Notarization using Blockchain for the FDRE Document Authentication and Registration Agency (DARA)" by Ashenafi Zena and Mesfin Kifle [3]

investigates the expediting of authenticating notarized documents using blockchain. The paper states that consumers of notarized documents (financial and civil service institutions) resort to authenticating notarized documents against registered copies at the notary as a measure of security. The researchers say such a lengthy security measure leads to delayed service delivery and customer dissatisfaction. Automating this key process has been slow for a long time due to the difficulty of creating efficient and effective digital trust relationship between transacting parties. In the paper, the researchers compare blockchain with similar technologies and adapt it for the purpose of notarization. Their proposed solution is a notarization, registration, and verification system based on an open source blockchain infrastructure. They hypothesized that, blockchain can deliver a secure, open and efficient solution to notarization. The researchers claim to show that development in distributed trust relationships with the advent of blockchain technology is a viable solution to many problems, such as, trusted third-party computation model, single point of failure and the related scalability and availability issues. The proposed solution promises instantaneous and independent verification for consumers of notarization services, providing an expedited service to end consumers, and immediate dissemination of authentication revocation, thereby closing a potential security vulnerability. In another case, academics from Ghana have written a paper [4] with the aim of exploring the potential application of blockchain technology in land acquisition and title registration processes. They say their paper contributes to the existing literature by formally documenting land acquisition procedures, title registration, and potential integration with blockchain technology. The study focuses on improving land acquisition and title registration process by leveraging on blockchain technology as a platform for recording and managing land transactions to achieve certainty of title and security of ownership. They say that as blockchain technology is a new platform, there is no available evidence of how such a novel technology could be leveraged for use in lands acquisition and title registration process in Ghana. [4]. Two other works, one from India [7] on “Securing Land Registration Using Blockchain” and another from the Republic of Georgia, “Integration of Blockchain Technology into a Land Registration System for Immutable Traceability – A Case Study of Georgia” published by Springer from the BPM (Business Process Management) 2019 Blockchain and CEE (Central and Eastern Europe) Forum Vienna, Austria and authored by Lazushvili et al says, “...critical issues for land are traceability of records, hazards of document forgery as well as vulnerability to various errors...”. The authors say that the case study has ascertained to them how blockchain technology resolves certain issues related to land management. [5].

What this paper tries to do is take the relevant points from the different conceptual discussions of the related works and attempt to make a prototype implementation for immutably storing of Addis Ababa’s publicly funded condominium housing project registration and ownership transfer data.

3. Exploratory and Comparative Analysis

The methodology used for this paper follows design science research process, particularly the process recommended by Takeda et al, as summarized by Offerman et al [2] in their paper entitled “Outline of a Design Science Research Process”. The process follows a phased approach, namely, the enumeration of problems at first (problem identification phase), followed by suggestion of possible solutions (solution design phase), a decision on a solution to be

adopted (evaluation phase), and then arriving at the production of an artefact as proof of concept.

The solution design phase involves exploratory literature research of available solutions and making a comparative analysis of the findings for suitability to the problem under study. Out of the dozens of blockchain platforms available for analysis, four have been selected for the purposes of the paper based on maturity, adaptability, and built-in features. They are Bitcoin, Ethereum, Hyperledger Fabric, and Multichain. Comparison metrics are used to assess the blockchain platforms selected for the study: Accessibility, Maturity, Performance / Scalability, Security, Consensus, and Relevance.

Accessibility metric refers to the “Platform’s Accessibility for Customization”, that is, the ease of programming the generic blockchain solution for particular use-cases. Maturity metric looks into “Maturity, Documentation, and Community Activity” for the platform under evaluation. It assesses how long the blockchain solution has been active in the industry, the variety of available documentation, and the community of practice. “Performance and Scalability” metric analyzes the transaction processing throughput and the network’s ability to adapt to operational capacity needs. “Security” metric refers to the privacy and confidentiality of each of the platforms. Consensus metric refers to the “Consensus Algorithm” or the mining method for the blockchain platform. It is a software algorithm that consists of an automated procedure through which all participants reach a common agreement about the present state of the blockchain network. Relevance metric weighs the “Relevance to the problem under study” in terms of the platform’s practicality in assuring trust for the public-funded housing project under case study.

4. The Proposed Solution

4.1 The Proposed Blockchain Platform

Table 1 Comparison of Blockchain Platforms

	Bitcoin	Ethereum	HyperLedger	Multichain
Accessibility / Customization	• GUI, CLI, API • Stack-based scripting	• GUI, CLI, API • High-level programming	• GUI, CLI, API • High-level programming	• GUI, CLI, API • In-built database
Maturity / Community	• The oldest • Rich documentation	• Second oldest • Ranks top for ‘Adoption’ empirically	• Benefits from the experience of Linux Foundation	• Inherited a mature codebase from Bitcoin
Performance / Scalability	• Limited performance and scalability	• Performs better and is scalable with PoS algorithm	• Performance / scalability addressed by design	• Performance / scalability addressed by design
Consensus Algorithm	• PoW	• PoW / PoS	• Plug-and-Play of choice	• PoW / Mining Diversity
Security	• Highly secured, never breached	• Breached once but rectified	• Operated by vetted participants	• Secured by diversity parameter
Relevance	• Costly, time taking for case	• Big and costly for case	• Oversized, complicated for prototype	• Adaptable for case

Narrative analysis put HyperLedger on the leading edge for enterprise level implementation of blockchain applications, this study recommends Multichain for implementation based on the

appropriateness strategy followed in the methodology while avoiding costly and complicated undertaking for a basic prototype implementation of the case study.

MultiChain deals with problems related to performance and scale by using its design features of being able to control the block size and other parameters of the blockchain.

MultiChain has two options to reach consensus, either PoW (Proof-of-Work) or Mining Diversity (round-robin type). Enterprises usually use the round-robin type since participants in a MultiChain network are known and that there is no need for an expensive consensus mechanism such as PoW. The basic premise is that as the participants within a private blockchain network, such as Multichain, are already “trusted” to some extent because they are identifiable entities, there is no need for the computationally expensive PoW algorithm to be used on such private networks. With this understanding amongst private blockchain participants (usually established businesses or other organizations), Mining Diversity protocol in Multichain can securely provide consensus to provide continuity to the chain.

MultiChain is secure enough as it allows users to configure parameters including the permissions to access the network, the privacy of the chain, the maximum block size, and the mining incentive. Multichain has a strong and active community of developers and a full-fledged documentation that is progressively updated. Its Github page is also rich with additional utility software for GUI applications and an online explorer.

As the main objective of this study is to help assure public trust with the use of blockchain technology in the conduct of government business in registering and allocating of houses, data is recorded permanently and immutably for the following items:

- Registrant name
- Date of registration
- Unique identification for each housing unit
- Results of lottery draws
- Date lottery was drawn
- Date house was transferred to registrant

It is expected that the prototype implementation will be a foundation for helping assure public trust in the project data that the system provides. The system is built in such a way that all the data on record is openly viewable and verifiable. The security and integrity of the system will be verified by checking the data against the existing, publicly available data of the project that has successively been published in rounds on the City Government’s official Gazette for more than a decade. The data on the Gazette has actually been the basis for the transfer of individual housing units to lucky winners of successive rounds of lottery draws.

The Trust Assurance system provides the public with an easy-to-remember website address to check all data related to registrants, lottery results, and transfer of houses to individuals with basic search functionality.

4.2 Data Structure Design

The Addis Ababa Houses Project data (records) can ideally fit into a large single table with no need for other related tables. The number of records and the data in each record grows as the project progresses and as new data is collected.

The storage and querying of this kind of massively growing data fits best in a key-value, non-relational (NoSQL) database. All of the data in such key-value databases consists of an indexed key and a value. Key/value databases are highly optimized for applications performing simple lookups using the value of the key, or by a range of keys. [6]

Every entity (record) in this kind of databases is a set of key/value pairs. A key has multiple components, specified as an ordered list. The major key identifies the record and consists of the leading components of the key. The subsequent components are called minor keys.

The main data fields (keys) for Addis Ababa Condominium Houses Project are:

- The housing unit identification
- The name of an eligible city resident
- The registration ID of an eligible city resident
- The date of registration
- The lottery round
- The date the housing unit was transferred to the lucky winner

From the data fields (keys) listed above, the mandatory one(s) for a record is/are “the housing unit identification” OR “the name, registration ID, and date of registration of an eligible city resident”. The other fields are optional and can be added as they become available. Additional fields may also be part of any record as needed, such as “Remarks” and “the date the lottery results were published on the official gazette”.

In most cases, the major key is “the housing unit identification” following the nomenclature, **NameOfSite_BuildingNo_FloorNoApartmentNo**, such as Kolfe_Bldg4_103 – Name of Site: Kolfe; Building No:4; Floor No: 1; Apartment No.:03

The minor keys may include:

- | | |
|----------------|---|
| • RegName | Registrant’s name |
| • RegID | Registrant’s unique ID of registration |
| • RegDate | Date of registration |
| • LottoRound | The round number of the lottery draw |
| • LottoDate | Date the lottery was drawn |
| • GazetteDate | Date the lottery result was published on the official paper |
| • TransferDate | Date the house was transferred to lucky registrant |
| • Remarks | Any comment as needed for the specific record |

For a new registrant who has not yet won the lottery, the major key for the record is the “registrant’s unique ID of registration” and the minor keys are the “registrant’s name” and the “date of registration.”

4.3 Proof of Concept (PoC) Demo

The system implements a peer-to-peer (P2P) distributed computing architecture where all nodes are equal. MultiChain Community Edition open source blockchain platform is selected for the prototype PoC demo, along with related utilities such as Multichain Command Line Interface for direct access to all APIs, Multichain WebDemo for the GUI interface, and Multichain Explorer for drilling down the data in finer detail. The in-built NoSQL non-

relational database of Multichain called “Stream” is used to store the key:value pair project data.

The PoC demo system is installed in the cloud on Amazon Web Services (AWS) EC2 (Elastic Compute Cloud) for ease of use, affordability, availability, and scalability. It is separately installed on two virtual machines (nodes) in line with the distributed and decentralized nature of the application with no single point of failure (SPOF).

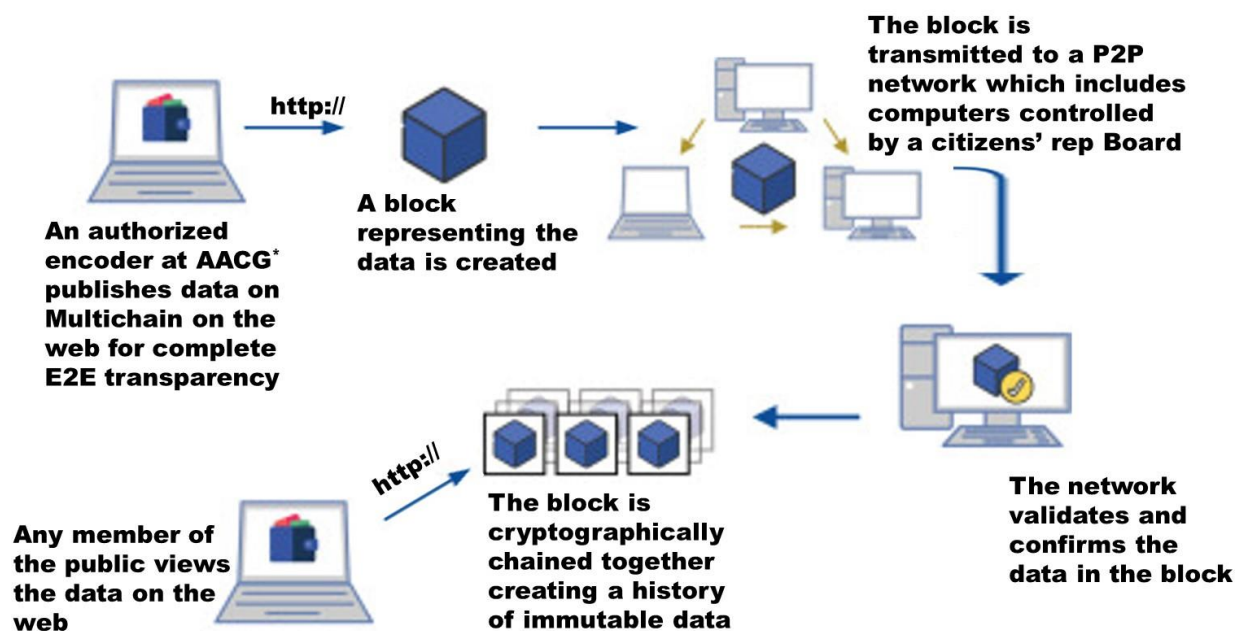
Amazon Linux is the underlying operating system, Apache is used for the web server, PHP for the front-end, and Python for some high-level scripting as recommended by Multichain.

Arbitrary representative data for the housing project is entered into the prototype PoC demo record by record. For large size implementation the data can be imported from a text file in JSON format.

In case of data entry error, an additional record has to be inserted with identical keys and indicating the error in the value of the “Remarks” key since any data once committed to the blockchain cannot be updated or deleted.

A super admin manages all permissions for the system, but cannot modify or delete any project data. A data uploader simply uploads publicly available project data into the blockchain system in a pre-defined JSON format.

The public is provided with a website address to view all data related to registrants, lottery results, and transfer of houses to individuals with some basic search functionality.



*Addis Ababa City Government

Figure 1: End-to-End Multichain Solution for Trust Assurance

The features below are objectively considered for the evaluation:

- Reliability – Accuracy of Search Results_: how the results of searches closely match anticipated outputs
- Versatility – Multiple-Key Selection : whether the prototype allows categorization of data with multiple major keys
- Security – Un-deletability of Data : whether there’s any possibility of removing data from the system once it’s entered
- Security – Un-modifiability / Immutability of Data : whether there’s any possibility of changing data values once they are entered into the system
- Auditability – Traceability of Data : whether the system provides ways to trace “who” entered a record and “when” exactly it was entered
- Chronology – Sequentiality (Keeping of Data Entry Chronology) : whether the system preserves the order of data entry without any possibility of rearrangement

Stream: AA_Cond_Houses – 2 of 2 items with key: Asko_Bldg12_401

Publishers	(1U24MaEHWxanjTJrzYEzRZn3DKssPnkGn1TB5R)
Key(s)	Asko_Bldg12_401
JSON data	{ "RegID": "7935", "RegName": "Tesfaye Alemu", "RegDate": "24-Oct-2006", "LottoRound": "9", "Remarks": "Lucky winner has not yet showed up; house put on hold" }
Added	2022-06-14 10:14:01 GMT (confirmed)

Publishers	(1U24MaEHWxanjTJrzYEzRZn3DKssPnkGn1TB5R)
Key(s)	Asko_Bldg12_401
JSON data	{ "RegID": "7935", "RegName": "Tesfaye Alemu", "RegDate": "24-Oct-2006", "LottoRound": "9", "Remarks": "Lucky winner is expected to start the process of ownership transfer" }
Added	2022-06-14 10:07:42 GMT (confirmed)

Figure 2: Querying by HouseID Returns the History of that House

5. Assessment Using the Evaluation Features

- Reliability – Accuracy of Search Results:

The search results returned exactly what was intended unless the search text was mistyped or not fully typed.

- Versatility – Multiple-Key Selection:

It was possible to use multiple major keys for data record categorization, such as HouseID and LottoRound.

- Security – Un-deletability of Data :

There was no way during the testing of the prototype to delete data once it was entered into the system, either from the back-end or front interface.

- Security – Un-modifiability / Immutability of Data :

There was no way during the testing of the prototype to change data values once they were entered into the system, either from the back-end or front interface.

- Auditability – Traceability of Data :

The prototype system readily provides the “Publisher” (the *who*) for every record and the unalterable system “timestamp” (the *when*) the record was published.

- Chronology – Sequentiality (Keeping of Data Entry Sequence) : In line with the system timestamp, the history of a certain house was shown from when it was ready for lottery draw until it was transferred to the lucky winner. There was no possibility of changing the sequence of events as and when they happened

6. Conclusions and Recommendations

The outcomes for the research questions are outlined below:

- The first question (RQ#1) asked whether blockchain technology can be used to record basic data about users registered for condominium houses, data about the housing units, and data about the results of successive rounds of lottery draws, and if so, whether the data thus recorded can be kept tamper-free, unchangeable, and stored for perpetuity. The PoC demo appropriately showed the answer is a clear “Yes”.

- The second question (RQ#2) asked whether it is possible to provide the immutable recorded information related to registered users, condominium houses, and corresponding lottery results accessible for public scrutiny. Again, the answer is “Yes” and the “how” has clearly been demonstrated by the PoC demo.

- The third question (RQ#3) asked how blockchain technology could help the public to trust the government-provided record of the housing project “beyond any reasonable doubt”. This is the core investigative quest of the thesis and answering it in the affirmative requires both technological solutions and public-awareness-creation approaches. A Citizen’s Board can play a crucial role in assuring the public that the technology offers an unforgeable platform for everyone to trust, which can easily and always be cross-checked with publicly available data and events since the inception of the housing project.

The following are recommended for future work.

- A full-fledged system for Trust Assurance for Addis Ababa Housing Project be implemented on an enterprise-grade blockchain platform, taking into account additional scenarios that are captured during the system analysis phase that involves representative stakeholders from the government and the residents of the City.

- The lottery allocation needs to be implemented using smart contracts on the enterprise-grade implementation, avoiding any manual interference.

- The enterprise implementation should include flexible ad-hoc queries that can go deep down into key:value pairs of minor keys for advanced querying functionality.

- It’s also recommended that the City Administration adopts the use of blockchain technology for internal audit purposes for regulatory and compliance inspections for the Housing Project.

References

- [1] W. Eggers, "Rebuilding Trust in Government -- Four signals that can help improve citizen trust and engagement," Deloitte, 9th March 2021. [Online]. Available: <https://www2.deloitte.com/us/en/insights/industry/public-sector/building-trust-in-government.html> . [Accessed 17th September 2021].
- [2] P. Offermann, "Outline of a Design Science Research Process," in *Proceedings of the 4th International Conference on Design Science Research in Information Systems and Technology*, Pennsylvania, 2009.
- [3] A. Zena, "Secure Notarization using Blockchain for the FDRE Document Authentication and Registration Agency (DARA)," *HiLCoE Journal of Computer Science and Technology*, p. 9, 2018.
- [4] K. Mintah, "Skin lands in Ghana and application of blockchain technology for acquisition and title registration," *Journal of Property Planning and Environmental Law*, 12 May 2020. [Online]. Available: [tps://www.emerald.com/insight/2514-9407.htm](https://www.emerald.com/insight/2514-9407.htm). [Accessed 27th November 2021].
- [5] Lazuashvili, "Integration of Blockchain Technology into a Land Registration System for Immutable Traceability – A Case Study of Georgia," in *BPM 2019 Blockchain and CEE Forum Vienna, Austria, September 1–6, 2019 Proceedings*, Vienna, Austria, 2019.
- [6] "What is NoSQL?," DataStax, 01 01 2022. [Online]. Available: [tps://www.datastax.com/what-is/nosql](https://www.datastax.com/what-is/nosql). [Accessed 08 06 2022].
- [7] Krishnapriya, "Securing Land Registration using Blockchain," 4th May 2020. [Online]. Available: [tps://www.sciencedirect.com/science/article/pii/S1877050920311649](https://www.sciencedirect.com/science/article/pii/S1877050920311649). [Accessed 22nd March 2021].
- [8] F. Zemenu, *Exploration of the Potentials of Blockchain in Helping Assure Public Trust in the Registration of Eligible Citizens and Transfer of Ownership of Public-Funded Houses in Ethiopia*, Addis Ababa: HiLCoE - Master of Science Thesis, 2024.

Image Based Similar Product Recommendation System For E-commerce Platform

Girum woldkidan

HiLCoE, School of Computer and Information Technology
ghmicky@gmail.com

Tibebe Beshah

School of information Science, AAU
tibebe.beshah@gmail.com

Abstract

Research on online product recommendations is crucial, as text-based search functionality has limitations. Image-based similarity search offers enhanced visual search experiences, benefiting e-commerce platforms. A machine-learning-based image-based recommender system was designed using Singular Value Decomposition (SVD) and Principal Component Analysis (PCA) to reduce dimensionality and enhance computational efficiency. The K-Means++ clustering method identified similar product groups, refining the recommendation process.

The study compared four unsupervised clustering algorithms on a dataset of 44,446 product images from the Kaggle platform. The PCA-SVD transformed K-means++ methodology showed higher performance, delivering higher recommendations compared to other strategies, indicating its potential for revolutionizing e-commerce product recommendations.

Keywords: K means-++;Principal component Analysis;Principal Component Analysis through singular value decomposition; Similar image recommendation

1. Introduction

Modern web services and the internet have been growing in the past few decades, making access to a wealth of information easier than ever[1]. In order to filter out the essentials, users may have a difficult time taking away information from this mass of information. Our inability to browse through all of the content on a website makes recommendation systems essential to our user experience, while also exposing us to more inventory we would otherwise never find. Recommender systems provide recommendations to users about the best option among a set of choices. The use of recommender systems is very useful in order to provide the user with the products or services he or she needs. The recommendation system is part of machine learning that helps users find products and content based on data on the website[2]. Recommendation systems have always been in demand in every industry and domain as matching a user's preference is always the most important thing in modern-day business[3].

In a product selling company, it is quite convenient to recommend products based on the user's needs in order to achieve a user-friendly environment. For this reason, e-commerce websites using recommendation systems to create a better internet shopping experience. Visual appearance plays a crucial role in marketing and selling clothing and electronic products. It has the ability to captivate consumers, enabling them to easily recognize and differentiate between different products, ultimately influencing their purchasing decisions. An image recommendation is concerned with finding related objects to a given object that is represented by an image. Image recommendation relies on extracting most descriptive and distinctive features and utilizing them to get matching recommendations [3]. In addition to suggesting relevant images based on the ones the user has already seen or liked, image recommendations can also suggest similar pictures. Using this feature can help users discover new images they might like.

It is highly crucial for an e-commerce platform to incorporate a sophisticated recommendation system that leverages image-based matching to not only drive sales but also provide users with a seamless experience. But most of E-commerce based search engines services employ text-based searching [4]. A textual or key-word-based recommendation system may misdirect users to inferior quality or related products that they do not need. In addition, this kind of approach had some limitations and missed many features such as colours, patterns, texture, and shape in the product images[5], [6].

Most of the solutions involved semi-image-based approaches or text-based searches, and those image-based approaches solution have low the performance and provide inaccurate recommendation. Most of these systems use textual meta-data of products such as attributes, descriptions, and purchase histories of the different users. This text-based search requires specifying the description of each product, which may or may not cover the entire item.

Hence, a picture of any product should clarify the user's expectations about its appearance, usage, and brand. In spite of the maturation of computer vision and image

processing, applications of image features for product retrieval using machine learning in online shopping are mostly still unexplored. In order to achieve this, it is also necessary to group similar items or images together with Clustering algorithms which use feature vectors extracted from images to group similar images together. It is essential and most challenging to select the right clustering algorithm for image classification dataset optimization.

To come up with a solution to those problems based on user interaction and interests, to search the products efficiently in an online shopping system using image-based searching techniques, which is giving an input image rather than text description. In other words, a user provides or insert an input image, and similar products (images) are presented to the user. we propose a machine learning based approach for a similar image recommendation system. objective of the study is to experiment and design an image-based recommendation system model for similar products on e-commerce websites that will help buyers find the alternative they're looking for.

2. Literature Review

Recommender systems filter information, providing personalized content, to reduce user effort and time. However, satisfaction decreases with more difficult selections, as noted by Iyengar et al. [7].

Recommender systems are algorithms that offer personalized recommendations based on user preferences. They involve users and items, who rate items based on their experiences. These systems can be content-based, collaborative, or hybrid, and their choice depends on the available data. Collaborative filtering is a suitable approach for historical interaction data. Content and context filtering techniques enhance recommendation accuracy by assessing user characteristics and item properties. These methods aim to predict user preferences and anticipate future interactions. The success of a recommender system depends on its ability to leverage diverse data sources and advanced algorithms to deliver personalized, relevant recommendations in real-time. Collaborative filtering is useful for past interactions.

➤ Content-based Recommendation System

A content-based recommendation system categorizes data items based on user preferences or past searches, enabling structured similarity judgments between products.

Salter et al.'s [8] study highlights the limitations of content-based recommendation systems, which primarily recommend items related to previously evaluated ones, limiting diversity in recommended content, despite their crucial role in user profiles.

Content-based recommendation systems use advanced technologies like text mining, Semantic Analysis, TF-IDF, Neural Network, Naive Bayes, and SVM to effectively derive user preferences. The MMP method reduces dimensionality in image retrieval, enhancing the margin between negative and positive examples.

➤ Collaborative recommendation

A collaborative recommender system uses user evaluations to generate recommendations based on similar interests. It constructs a user's preference database and predicts items that match a user's tastes. By starting with a user, it finds similar users in neighbourhoods, and recommends items to the user. The efficiency of a collaborative algorithm depends on its accuracy in finding the target user's preferences and interests.

A collaborative recommender system uses user evaluations to create a preference database, predicting items that match a user's preferences, as per Miller et al. and Hars et al.[9].

There are two types of collaborative recommendation system: memory-based approach and model-based approach [8]. Memory-based collaborative approach suggests new items based on surrounding preferences, using technologies like Pearson correlation, vector cosine correlation, and KNN to create similar groups among users. It can be further segmented into User-Based Collaborative recommendation and Item-Based

Collaborative recommendation systems. Model-based collaborative approach predicts user ratings for unrated items using data mining and machine learning algorithms, extracting features from the dataset.

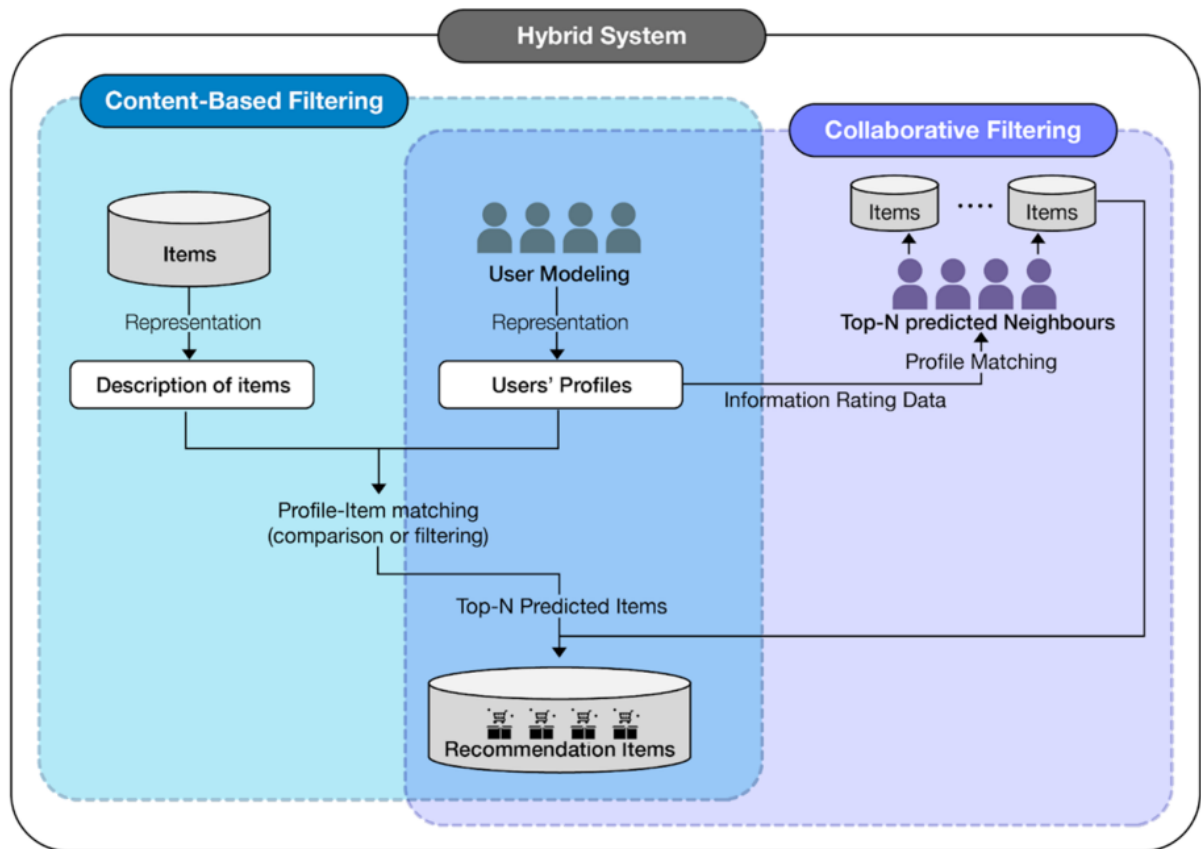


Figure 1 process of general recommendation system model[10].

With respect to empirical works, there are a couple of attempts that considered content and collaborative recommendations. Younus et al. [11] proposed a content-based image retrieval system using PSO and K-means, comparing it to existing methods. The system performed better than existing methods, based on histograms, co-occurrence matrix, color, and wavelet moments.

Facultyrrurozi et al.[12]developed a fruit query system using fuzzy-based feature extraction, color histogram, and K-means clustering. They tested their model on 1075 images, achieving 89.5% accuracy and 90% accuracy for single and multi-object fruits.

A road-based travel recommendation system is developed in [13]using geo-tagged images. The proposed approach recommends the most famous landmarks and travel routes. The proposed algorithm is implemented and tested with Flickr photos in Munich, Germany. Totally, 45,950 images were acquired via Flickr API (<http://www.flickr.com/services/api/>). And they filter out unnecessary images whose tags are not related to tourism. Consequently, 31,335 geotagged images taken by 2422 users were finally chosen as the study data in this paper. the total test accuracies of these study were rather good (approximately 80%).

Farz. Jouyandeh, P. Zadeh et al. [14] developed an Image-based Personalized Advertisement Recommendation System using social media analysis. They evaluated the system using Flickr 8K dataset and Twitter API, finding that 74.166% of selected users were related to associated images with a score above 0.50.

The study by Gan et al. [15] developed StyleNet, a trainable framework for creating captions for photos and videos using the FlickrStyle10K dataset. The framework, which combines Deep Learning techniques and statistical methods, was favoured by over 80% of judges.

Karthikeyan and Aruna [16] developed a probabilistic text and image-based semi-supervised clustering approach using 2000 documents from five categories. Their method outperformed two existing algorithms and outperformed CBIR (content-based image retrieval) and signature-based approaches. They also compared their method to K-means and Dbscan.

3. Methods and Materials

The study employs quantitative experimental research, using unsupervised machine learning techniques like PCA-SVD for dimensionality reduction and K-means ++ for clustering and distance measures between top-K product image recommendations. This study uses data from Kaggle, an online community platform for data scientists and machine learning enthusiasts, to analyse product images. The dataset includes 44446 high-resolution images with ten attributes and two additional attributes. The images are 25 GB in size and include various product categories and 45 subcategories, including apparel, accessories, footwear, personal care, and sporting goods. As with the pre-processing data stage, seven thousand three hundred thirty-three 7333 images are sampled out using Random sampling technique.

The experiment involves training and testing anaconda python 3.12 on an Acer Swift 3 laptop with a 11th Gen Intel Core i7-1165G7 @2.80GHZ with 8.00GB RAM and 500 GB hard disk and Microsoft Windows 10 64-bit operating system. Data is collected and processed using Python libraries like NumPy, pandas, matplotlib, and scikit-learn. The initial step is to compute PCA singular values and compute the K-means++ clustering approach for the sampled dataset. The paper also calculates the cluster's inertia and distance measure using Manhattan distance measure. The experiment includes 45 subcategories of various product images.

The model is trained on training and testing images using feature engineering and machine learning techniques like Principal Component Analysis, SVD, similar distance measure, and Probabilistic PCA. The best recommendations are determined using distance from the origin to the database.

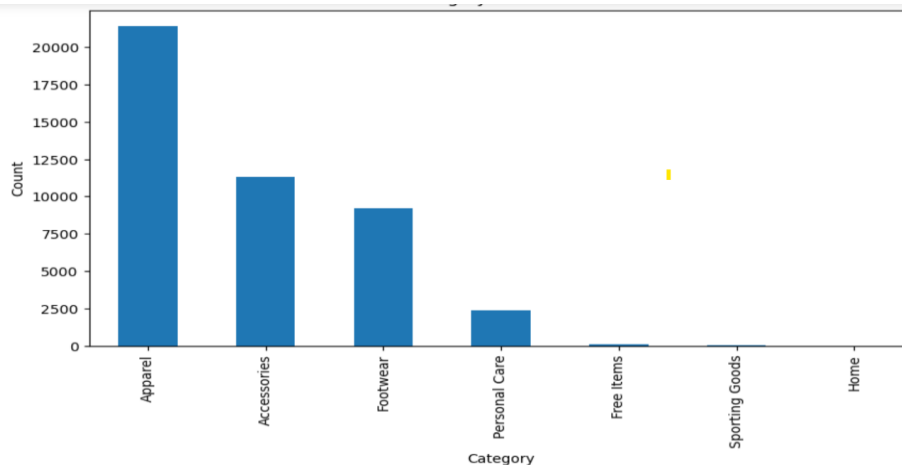


Figure 2 Major category distribution.

4. Dataset and Experimentation

Techniques like data cleaning, normalization, balancing, and numericalization are used to prepare data for analysis. These techniques are used to map symbolic and non-numerical values to numerical values, ensuring data is semi-prepared for analysis. According to this removing incorrect, incomplete, irrelevant, duplicated, or improperly formatted data. In this research, three irrelevant data values are dropped from a Data Frame with 12 attributes. machine learning algorithms to function properly, only numerical values are required as inputs or matrix values should be input values for machine learning algorithms. Therefore, the non-numeric features, such as 'gender', 'masterCategory', 'subCategory', 'articleType', 'baseColour', 'season', 'usage' and 'productDisplayName' features are converted into numeric values using a One-hot encoding scheme, which converts each category into a binary column and returns a sparse matrix or dense array. Once the dataset features are numerically analyzed, they have different values in different ranges/scales. In this case, the collected datasets failed to yield appropriate results, and the learning algorithm was unable to be effective. Therefore, a data normalization technique is applied to transform dataset features into a common range, so that larger numeric feature values do not dominate smaller numeric feature values. After all these processes we seat list of the configuration setup to be carried out for the experiments.

Table 1 *List of experimental settings to be conducted*

Number of experiments	Algorithm to be used	Data to be used	Method for finding optimal K value	Cluster distance measurement	Performance metrics to be used
	PCA_SVD to be computed first				
Experiment 1	K-means ++	PSVD transformed image	Elbow with 8-16 Range of K value	Manhattan	SC index, CH index and DB index
Experiment 2	K Medoids				
Experiment 3	DBSCAN				
Experiment 4	Agglomerative Clustering				

The experiment begins with applying the PSVD method for dimensionality reduction to identify determinant features and principal components. A 92% explained variance threshold is set, and PCA singular values are calculated on a sampled dataset and resulting in 2900 final principal components.in other word the PCA singular values were calculated using the PSVD process on a sampled dataset, resulting in 2900 final principal components after reducing the dimensionality to 92.01%. The elbow method determines the optimal K value for clustering algorithms by assigning a range of values to k,

resulting in the optimal K value of 12. This process is followed by the four clustering algorithm using these optimal k values. By comparing data points based upon their similarity.

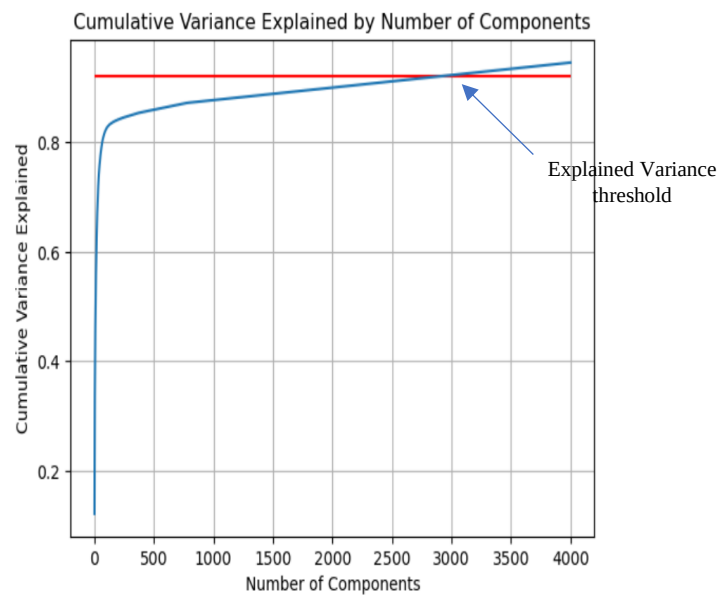


Figure 3 Cumulative Variance

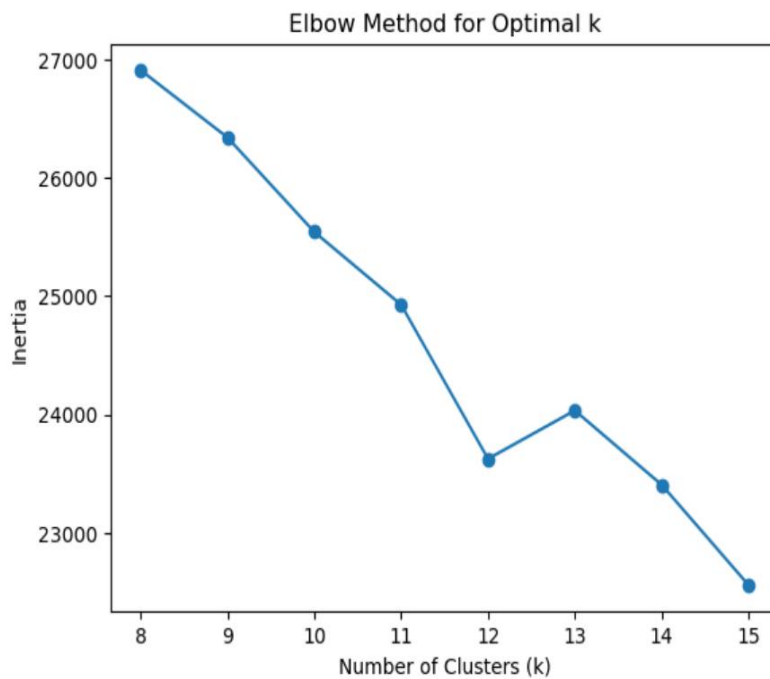


Figure 4 Elbow Method for finding optimal K value.

5. Result and Discussion

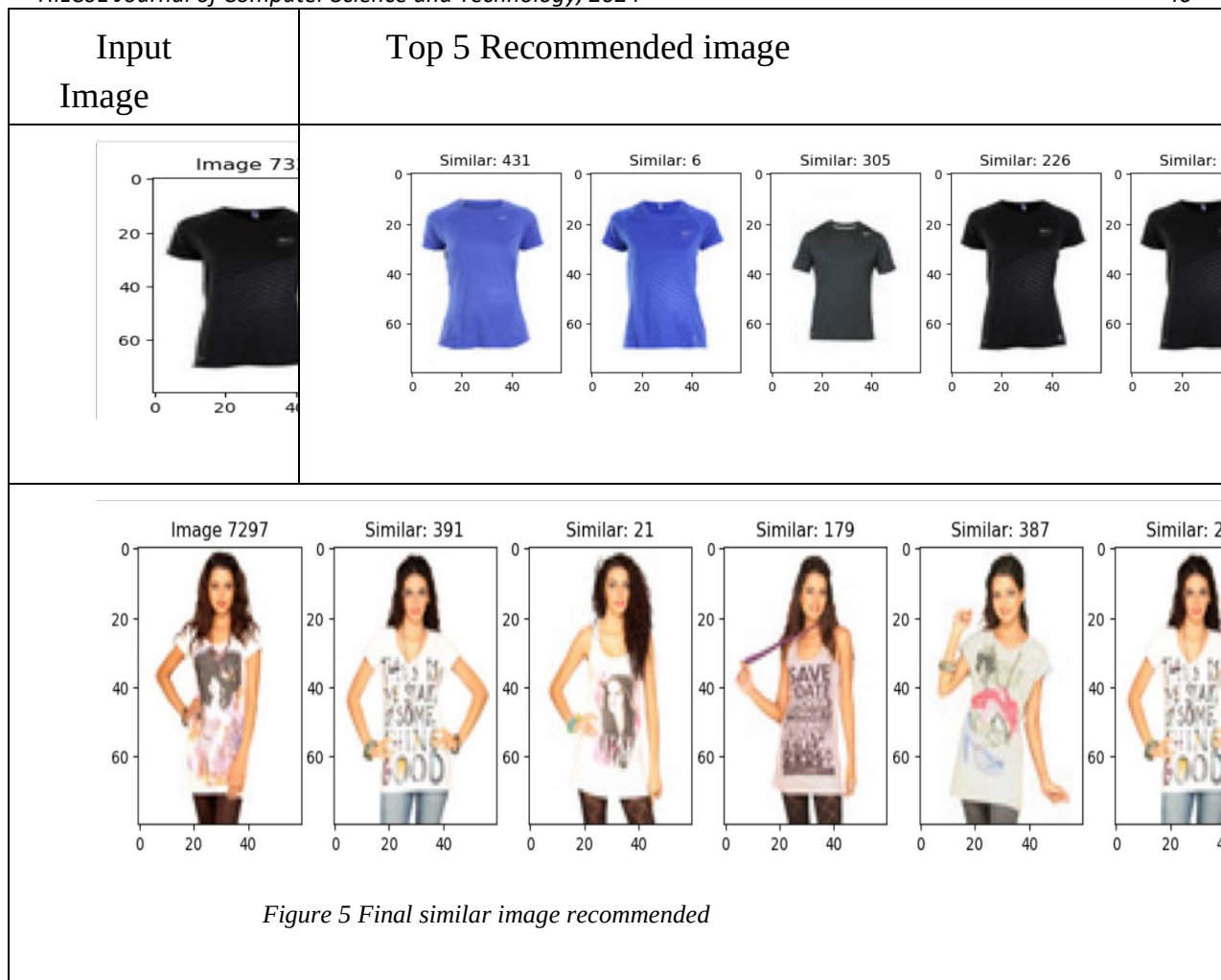
The experimental results show that using four types of clustering approaches on PSVD-transformed images of 2900 final principal components significantly reduces the size of the dataset. This research analyzed cluster performance metrics to improve recommendations for unsupervised learning techniques. Cluster performance metrics are crucial for evaluating the quality of clustering algorithms, as they help identify distinct similar clusters within a dataset. Cluster evaluation measures homogeneity and separation between groups using metrics like the Silhouette Score. It helps identify hidden patterns in data for optimization.

The experiment computed several cluster performance metrics, including the Silhouette coefficient (SC), Calinski-Harabaz Index (CH), and Davies-Bouldin Index (DB). The k-means++ clustering algorithm performed better than the other three, demonstrating better separation between clusters and shorter execution time. These metrics, along with the Calinski-Harabasz index and Silhouette coefficient, contribute to better clustering quality.

Table 2 *Performance comparison of clustering algorithms using PSVD transformed image.*

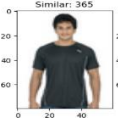
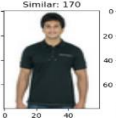
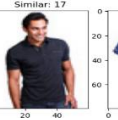
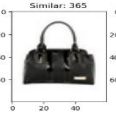
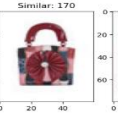
Type of Algorithm	Cluster performance metrics			Execution time in seconds.
	SC index	CH index	DB index	
K-means ++	0.1377	436.6845	2.2896	0.4104
DBSCAN	0.0893	292.3833	2.1351	159.6118
K medoid	-0.0253	8.8513	27.1432	1.9689
Agglomerative Clustering	0.1267	425.6170	2.3566	46.3894

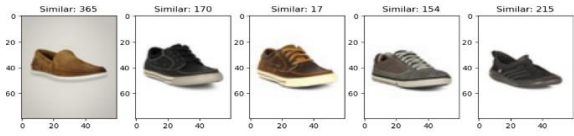
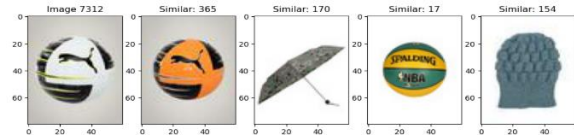
This experiment uses Manhattan distance to measure distance between clustered images in a grid or vector space. Top similar images are selected using K-means ++ clustering, and the absolute difference between coordinates is calculated. Based on Manhattan distance results, the top 5 similar images are recommended, providing a comprehensive overview of similarity analysis results.



The system's effectiveness should be thoroughly evaluated using various methods. A test case was used to verify the solution's workability. Clustering performance evaluation techniques were used to measure cluster separation and similarity. The system's precision and correctness were assessed using various test cases. MasterCategory values were used as a test case to ensure accurate performance.

Table 3 System Performance Evaluation

Test case		Expected output	Actual output	Remark
master Category	Article Type			
Apparel	T-shirts	Top 5 similar T-shirts	    	Accurate performance "5/5"
Accessories	Handbags	Top 5 similar Handbags	    	Poor Performance "1/5"

Footwear	Casual Shoes	Top 5 similar Casual Shoes		Accurate Performance "5/5"
Sporting Goods	Basketballs	Top 5 similar Basketballs		Poor Performance "2/5"

This study investigates clustering algorithms to develop a product recommendation model for an e-commerce platform based on images. The goal is to provide users with knowledgeable product recommendations, aiding their decision-making process. Nine features were tested, with two being more determinate: master category and subcategory. Poor performance results are evaluated using masterCategory values as a test case, ensuring accurate performance. It has been shown in other studies that their image-based recommendation system uses three features: texture, color, and shape[4], [17][18].The research identifies k-means++ clustering as the most suitable algorithm for accurately identifying similar images, as demonstrated in experiments comparing four methods. The k-means++ clustering method was used for recommending systems in [19], [20]and showed better performance.

6. Conclusion

The research aims to simplify searching for products on an e-commerce platform by suggesting similar images using Kaggle datasets, focusing on bringing the desired product to the customer. The experiment uses unsupervised machine learning to produce similar product images for e-commerce platforms. It uses 44446 images from Kaggle and uses K-means++ clustering algorithms. Anaconda Jupiter notebooks simulate experiments, and accuracy and execution time are calculated using clustering performance metrics. The study used PSVD dimensionality reduction on a product image dataset, achieving a 92.00% variance. The images were then clustered using K-means++. Compared to other clustering algorithms, PSVD-K-means++ scored well on similarity coefficients and CH scores, but lacked in DB index score. This suggests that similar image recommendation algorithms work well. The study identifies determinant features for optimizing item search in ecommerce systems, including gender, masterCategory, subcategory, articleType, baseColour, season, year, and usage attributes. MasterCategory and subcategory are the most significant determinant variables. K-means ++ is found effective.

The results obtained from this experiment can be considered a very effective result. However, a lot of work can be done to improve performance of the proposed system. In order to improve the effectiveness of image-based recommenders, continuous experimentation and research are essential. For more accurate image feature extraction and training for similar image recommendations, it is suggested that using image data augmentation techniques and deep learning approaches (CNNs) is better

Reference

- [1] J. Beel, S. Langer, M. Genzmehr, B. Gipp, C. Breiting, and A. Nürnberger, "Research paper recommender system evaluation: A quantitative literature survey," in *ACM International Conference Proceeding Series*, 2013. doi: 10.1145/2532508.2532512.
- [2] Badreesh Shetty, "An In-Depth Guide to How Recommender Systems Work", Accessed: Sep. 26, 2023. [Online]. Available: <https://builtin.com/data-science/recommender-systems>
- [3] Zakim Medium, "Image Based Product Recommendation System," May 2020, Accessed: Sep. 26, 2023. [Online]. Available: <https://zakim.medium.com/image-based-product-recommendation-e1bfa29e508>
- [4] F. Ullah, B. Zhang, and R. U. Khan, "Image-Based Service Recommendation System: A JPEG-Coefficient RFs Approach," *IEEE Access*, vol. 8, no. D1, pp. 3308–3318, 2020, doi: 10.1109/ACCESS.2019.2962315.
- [5] M. S. Pal and S. K. Garg, "Image Retrieval : A Literature Review," *Int. J. Adv. Res. Comput. Eng. Technol.*, vol. 2, no. 6, 2013.
- [6] UKEssays. com, "Limitations Of Text Based Image Retrieval Psychology Essay," UKEssays. Accessed: Sep. 28, 2023. [Online]. Available: <https://www.ukessays.com/essays/psychology/limitations-of-text-based-image-retrieval-psychology-essay.php>
- [7] S. S. Iyengar and M. R. Lepper, "When choice is demotivating: Can one desire too much of a good thing?," *J. Pers. Soc. Psychol.*, vol. 79, no. 6, 2000, doi: 10.1037/0022-3514.79.6.995.
- [8] S. H. Park and S. P. Han, "Empirical analysis of the impact of product diversity on long-term performance of recommender systems," in *ACM International Conference Proceeding Series*, 2012. doi: 10.1145/2346536.2346592.
- [9] I. Im and A. Hars, "Does a one-size recommendation system fit all? the effectiveness of collaborative filtering based recommendation systems across different domains and search modes," *ACM Trans. Inf. Syst.*, vol. 26, no. 1, 2007, doi: 10.1145/1292591.1292595.
- [10] B. Walek and V. Fojtik, "A hybrid recommender system for recommending relevant movies using an expert system," *Expert Syst. Appl.*, vol. 158, 2020, doi: 10.1016/j.eswa.2020.113452.
- [11] Z. S. Younus *et al.*, "Content-based image retrieval using PSO and k-means clustering algorithm," *Arab. J. Geosci.*, vol. 8, no. 8, 2015, doi: 10.1007/s12517-014-1584-7.
- [12] Erwin, M. Fachrurrozi, A. Fiqih, B. R. Saputra, R. Algani, and A. Primanita, "Content based image retrieval for multi-objects fruits recognition using k-means and k-nearest neighbor," in *Proceedings of 2017 International Conference on Data and Software Engineering, ICoDSE 2017*, 2018. doi: 10.1109/ICODSE.2017.8285855.
- [13] Y. Sun, H. Fan, M. Bakillah, and A. Zipf, "Road-based travel recommendation using

geo-tagged images,” *Comput. Environ. Urban Syst.*, vol. 53, 2015, doi: 10.1016/j.compenvurbsys.2013.07.006.

[14] F. Jouyandeh and P. M. Zadeh, “IPARS: An Image-based Personalized Advertisement Recommendation System on Social Networks,” in *Procedia Computer Science*, 2022. doi: 10.1016/j.procs.2022.03.050.

[15] C. Gan, Z. Gan, X. He, J. Gao, and L. Deng, “StyleNet: Generating attractive visual captions with styles,” in *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, 2017. doi: 10.1109/CVPR.2017.108.

[16] M. Karthikeyan and P. Aruna, “Probability based document clustering and image clustering using content-based image retrieval,” *Appl. Soft Comput. J.*, vol. 13, no. 2, 2013, doi: 10.1016/j.asoc.2012.09.013.

[17] J. Zhou, X. Liu, W. Liu, and J. Gan, “Image retrieval based on effective feature extraction and diffusion process,” *Multimed. Tools Appl.*, vol. 78, no. 5, 2019, doi: 10.1007/s11042-018-6192-1.

[18] H. Tuinhof, C. Pirker, and M. Haltmeier, “Image-based fashion product recommendation with deep learning,” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2019. doi: 10.1007/978-3-030-13709-0_40.

[19] P. Phorasim and L. Yu, “Movies recommendation system using collaborative filtering and k-means,” *Int. J. Adv. Comput. Res.*, vol. 7, no. 29, 2017, doi: 10.19101/IJACR.2017.729004.

[20] M. Garanayak, S. N. Mohanty, A. K. Jagadev, and S. Sahoo, “Recommender system using item based collaborative filtering (CF) and K-means,” *Int. J. Knowledge-Based Intell. Eng. Syst.*, vol. 23, no. 2, 2019, doi: 10.3233/KES-190402.

Challenges and Readiness for Cybersecurity Threats at Selected Private and Governmental Banks

Betelhem Kidane Lema

HiLCoE School of Computer Science and Technology, Addis Ababa, Ethiopia

bent.nebo@gmail.com,

Tibebe Beshah

HiLCoE School of Computer Science and Technology, Addis Ababa, Ethiopia

tibebe.beshah@gmail.com

Abstract

The risk level from cybersecurity threats faced by banks has surged in recent years due to the global increase in cyberattacks, with banks being prime targets. Sensitive data and organizational security are at significant risk due to these escalating threats. This situation is exacerbated by the absence of a standardized cybersecurity framework to safeguard online financial data transfers between banks and their clients. To address these challenges, a new cybersecurity framework has been developed, serving as a guide for financial institutions to protect their data effectively.

To meet the research objectives, a mixed-method approach, incorporating both quantitative and qualitative methods, was employed. The survey questionnaire and interview outline were created by partially adapting existing and developing new ones as needed. Cybersecurity professionals from the Commercial works Bank of Ethiopia, Development Bank of Ethiopia, Nib International Bank, and Tsedey Bank participated in this study. The quantitative component involved 50 respondents from both private and government banks, representing a variety of job titles, responsibilities, and work histories. For the qualitative study, direct interviews were conducted with four bank cybersecurity directors. Alternatively, the researcher could use the census approach, also known as the complete enumeration survey method. Quantitative data were analyzed using the SPSS software program and relevant descriptive statistical procedures, while open coding techniques were used for the qualitative data analysis.

The study results indicated that social engineering attacks (59.1% in private banks and 71.5% in government banks), phishing attacks (45.4% in private banks and 46.4% in government banks), and DDoS attacks (32.1% in private banks and 22.7% in government banks) are the most prevalent threats facing the banking sector. Malware also poses a significant challenge.

Keywords: cybersecurity, cybersecurity framework, cybersecurity challenge and Cybersecurity readiness.

1. Introduction

Information technology is crucial for the success of financial organizations, offering efficient methods to collect, store, retrieve, process, transmit, and distribute data. It is seen as a competitive edge in today's work environment. The financial sector has greatly benefited from advancements in information technology, becoming increasingly reliant on it. As the volume of sensitive digital data grows rapidly, the impact of data breaches on individuals and systems also intensifies.

The banking industry, in particular, has been significantly influenced by the rapid growth of the information technology sector, making it a frequent target for IT-based attacks. Consequently, cybersecurity has become one of the primary concerns for banks.

As stated by Cisco [3], cybersecurity involves safeguarding systems, networks, and software from online threats.

Modern banking in Ethiopia began in 1905 and has seen continuous advancements up to the present. The integration of information technology services such as ATMs, Internet Banking, Mobile Banking, POS, and SMS has significantly enhanced the range of financial services available. Similar to other developed nations, Ethiopia is increasingly focusing on cybersecurity.

Ethiopian financial institutions are frequently targeted by cyberattacks, mirroring global trends. An INSA report from May 2022 highlighted an unprecedented surge in cyberattack attempts within a six-month period, predominantly targeting the banking sector. The report indicated that the Commercial Bank of Ethiopia accounted for 51% of the money lost to fraud, followed by the Bank of Abyssinia (17%), Oromia Bank (8.5%), and Wegagen Bank (8.2%). The threat landscape for financial institutions continues to grow.

According to the researcher's knowledge, there is scarcity of empirical studies in our country that assessed how prepared financial institutions are for brand-new cybersecurity challenges. Though there are some attempts there still is a need for Up-to-date empirical research to address an ever-changing threats and challenges. Therefore, this study aims to fill the research gap besides comparing the challenges posed by threats to the private and government banking sectors in Ethiopia, evaluating their readiness, Identifying the measures used by financial institutions to tackle the cybersecurity challenge, proposing ways to address recently emerging cybersecurity challenges, and developing cybersecurity a framework for the financial sector.

To address cybersecurity challenges, organizations need to pinpoint security issues to protect the confidentiality, availability, and integrity of their systems, thereby preventing asset losses from cyber-attacks. This study evaluates the cybersecurity threats faced by private and governmental institutions and examines how prepared these organizations are to tackle these challenges.

Thus, this study attempts to investigate and answer the following research questions:

- What are the challenges of cybersecurity threats on Ethiopian financial institutions?
- What readiness do Ethiopia's private and governmental financial institutions have for the challenge of cybersecurity?

2. Related Work

Agility (2022) states that the banking industry implements various measures, such as firewalls, intrusion detection systems, and encryption technologies, to safeguard their online systems and networks from cyber-attacks. The African financial sector is also vulnerable to cyber threats and incurs significant losses. However, there is a lack of tools to effectively collect, analyze, consolidate, and share data on a regional level. Consequently, breaches and losses in small or medium institutions often go unreported, as highlighted by the 11th annual EY/IIF global bank risk management survey.

According to [30], information security includes all the methods and devices enterprises use to protect data. This involves policies that prevent unauthorized access to company or personal data. The field of information security covers various topics such as network and infrastructure security, testing, and auditing, and is continually evolving. Information security aims to prevent unauthorized access, modification, or recording of sensitive data, as well as any other forms of disruption. It ensures the privacy and protection of critical data, including financial information, intellectual property, and client account details.

This research delves into recent studies on cybersecurity challenges. The first study discussed is “Assessment on Challenging Threats of Cybersecurity and its Emerging Trends on Selected Ethiopian Banking” by Bayu [8]. This article investigates the significant cybersecurity threats and emerging trends in selected financial institutions. It highlights the most prevalent cybersecurity threats in the banking sector and notes the absence of control and protective measures to mitigate these risks. Based on the study's findings, a cybersecurity framework has been proposed to address these challenges, along with recommendations for the selected banks.

The research titled “Internet Banking Security Framework in the Case of Ethiopian Banking Industry” by Aychiluhim [13] is relevant to this study. As internet banking becomes increasingly prevalent in Ethiopian banking, cybersecurity emerges as a significant risk. The study aims to outline internet banking security and proposes a baseline architecture for the Ethiopian banking sector. It concludes that for Ethiopia to achieve the most secure internet banking channel, banks must adopt a comprehensive and integrated approach. This includes involving all key stakeholders, such as banks, consumers, authorities, and service providers, to adhere to the recommended security framework. The paper ultimately advocates for a holistic, multi-layered security approach that extends to client-side security and incorporates national financial and security intelligence teams.

Mohammed [18] conducted a study on cyber-attacks and cybersecurity readiness in Iraqi private banks. The study aimed to assess the cybersecurity preparedness of the private banking sector. Using a case study methodology and quantitative techniques, the researchers collected data primarily through a questionnaire designed based on a literature review and expert recommendations. The study did not develop a framework for cybersecurity readiness. It found that while consumers occasionally face cyber-attacks and remain concerned about them, Iraqi private banks are actively working to enhance their cybersecurity systems. The study did not suggest any further research.

The previous literature review indicates that extensive research has been conducted on related topics. Most researchers have focused on cybersecurity challenges, risks, and solutions by reviewing existing literature, without specifically examining the challenges faced by the financial sector. Only one researcher has investigated the cybersecurity threats within a single banking industry [8]. This study addresses a gap by assessing and investigating cybersecurity threats and readiness through questionnaires, in-depth interviews with bank security experts, and a review of the bank's security policy. However, there has not been a comprehensive study on the cybersecurity threats and readiness levels comparing private and governmental banking industries in Ethiopia.

3. The Proposed Solution

This paper tried to overcome cyber security issues that are a major threat for organizations information security by proposing frameworks. The proposed cyber security framework will assist the financial institutions in terms of creating cybersecurity awareness and good practices to its employees to strengthen security by mitigating vulnerabilities for attacks. The key findings are related with the two categories: cybersecurity threat challenges and the banks preparedness in tackling cybercrimes.

In the first category phishing, DDOS and social engineering are the top cyber-attacks in both government and private banks. In "email security" employees' knowledge to phishing attack links is not good. In "safely use of internet and computer" personal computers security precautions are not practiced good enough, password management is also poor. In "threats and preventive measures" respondent security practices are below average. For instance, updating software, antivirus usage, security scanning and backup are not performed regularly in a timely manner. In "password management and security" respondents' password setting, changing and protection practices are unsatisfactory. In "information security terms and social engineering" section respondents' knowledge to information security responsibility and social engineering needs improvement. In "security standards, procedures and training" section the organization doesn't use information security standards, doesn't has well-organized procedures in case of cyber-attacks and employees are not being trained of information security. In "firewall, IPS, penetration and traffic control" the organization is not performing internal and external penetration testing. The implementation and usage of IPS/IDS and web filtering is unsatisfactory. In "wireless network security" the organization wireless network implementation and security management are not good. In the rest six sections multiple

security features are not implemented enough to strengthen the organization security posture such as port-based filtering, enabling port security, disabling unused ports etc. Ignoring these important issues may cause vulnerabilities to multiple cyber-attacks such as denial of service attacks and a business down time which is unacceptable to a financial institution. Overall the employees' information security awareness is not satisfactory and needs to be dealt with such kind of programs to protect its assets or even its business. Open line communication between financial institutions is also poor hence knowledge base systems is mandatory. Any platform either from information security administration (INSA) or national bank is encouraged.

This study is proposed in order to study and identify the challenges and preparedness in private and government banks. To secure the financial institutions assets there is a need for appropriate framework for organizations based on their context. NIST framework (2017) stated that the framework is not a one-size-fit all approach to managing Cybersecurity risks for information assets.

According to Hevner (2004), in order to validate the research, design science research relies on a thorough examination and re-construction of the design artifact contributions as described in the fourth principle. The suggested framework is examined in two rounds even if it is not thoroughly tested. With the chosen cyber specialists, a focus group discussion served as the initial phase of expert validation. The initial draft framework's completeness, correctness, and clarity were to be discussed, along with any room for improvement, during the focus group discussion. Individual conversations with the experts were then held to have a thorough discussion and get more comments. The proposed framework was updated to increase its completeness, clarity, and accuracy in response to the experts' feedback.

After the proposed framework was changed, a second round of expert validation was carried out using the survey to collect individual opinions on the revised framework's completeness, clarity, and correctness. During this stage, additional IT division staff members were invited to take part in the survey by briefly outlining the suggested framework.

The Hevener (2004) evaluation criteria, which include fit to the organization, comprehensiveness, reliability, clarity, correctness, and usability quality attributes, served as the basis for the questionnaire's content.

Finally, the abellow cybersecurity framework was developed by associating the findings of the research associating with NIST frameworks.

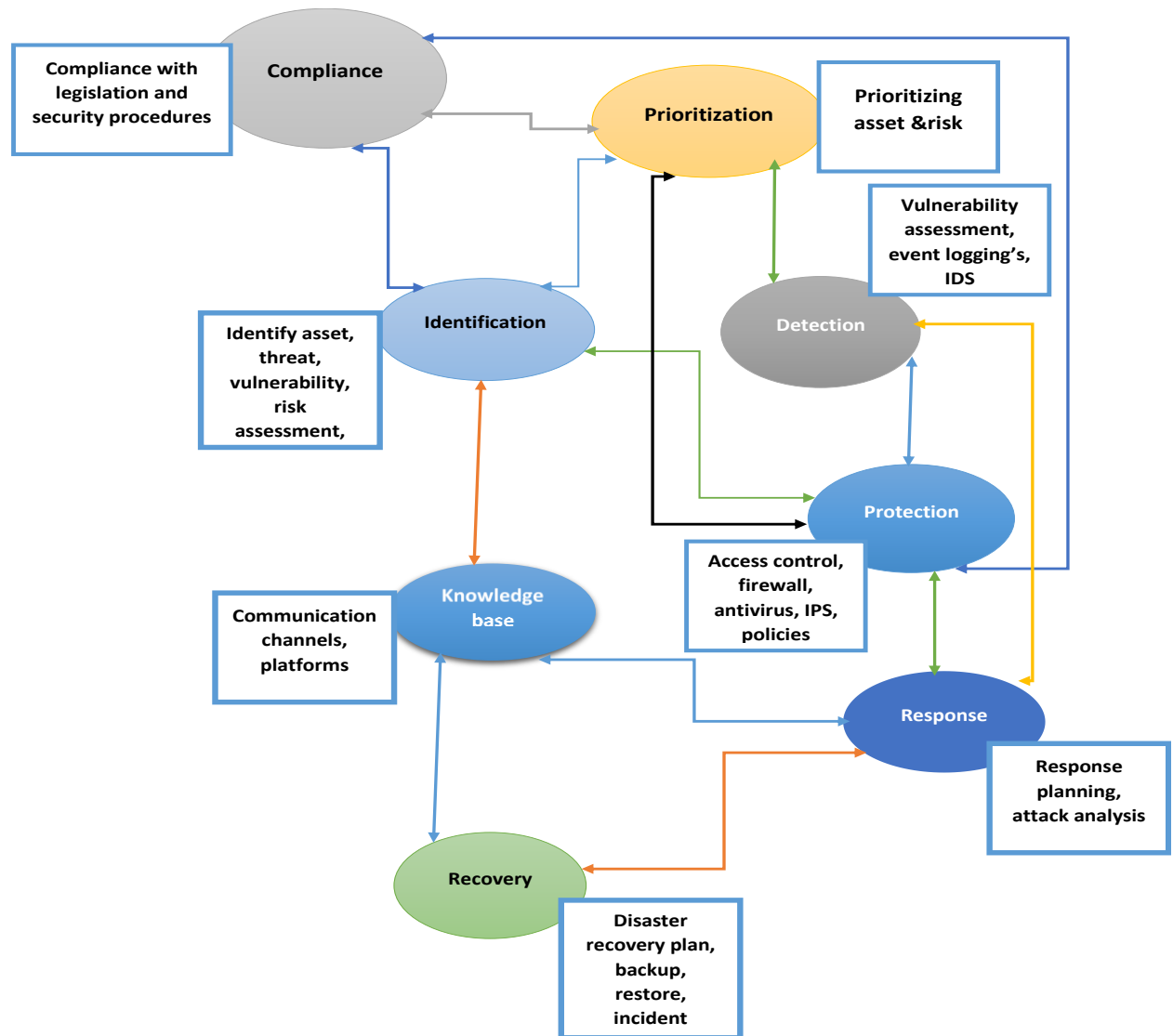


Figure 1. The proposed cybersecurity framework

4. Conclusion and Future Work

In order to customize a cybersecurity framework, this research examines cybersecurity challenges and readiness at private and government banks. Based on this research, the findings showed that social engineering, phishing and Distributed denial of service (DDOS) attacks are mentioned as the most security challenge in both private and government financial institutions. On the other hand, session high jacking attack is less visible too. Hence, special attention should be given on the three types of cyber-attacks. In order to alleviate and combat security issues, classification of those challenges should be important. Topics related to potential cybersecurity challenge, occurrence of cybersecurity challenge, cybersecurity policy and procedures, man power and capacity building, banks cooperation with other financial institutions and availability and use of cybersecurity infrastructures solutions are deeply discussed. In order to identify the challenges and difficulties of cybersecurity at selected private and government banks,

framework was developed in combination with NIST framework Core process, specifically identify, detect, prevent, respond, and recover.

The study's analysis and results led to the following five conclusions, which are listed below:

- Lack of well-prepared vital infrastructures like web gateway and filtering solutions; disaster recovery sites, email filtering gateway is clearly seen.
- Lack of internal experience, bad leadership skill, lack of dedicated security teams are the issues related with challenges existed. Similar issues with inadequate funding, poor open line communications contribute to the severe security issues.
- The study found that attacks like phishing, DDOS and social engineering the most security challenge of both private and government banks.
- The study found that the challenges lacked effective defenses against insider risks since most of the cybersecurity issues arises from insider.
- Although there were challenges in the financial institutions, the banks are at the high preparedness level in order to combat the growing risk of cybersecurity.

The research's results and conclusions highlight the issues, growing trends, and level of preparation for cybersecurity as well as any important cyber threats or exposure.

This research proposed a conceptualized framework that is used to guide the financial institutions to safeguard their information assets. The future scholar wider the scope to incorporate all Government and private institutions to have a generic cyber security framework Considering new technologies and infrastructures. In addition, the researcher will consider developing a knowledge base system for cyber security solutions.

References

- [1] Selgin and G. A, "Encyclopedia Britannica," 6 April 2022. [Online]. Available: <https://www.britannica.com/topic/bank>.
- [2] Afi, "CYBERSECURITY FOR FINANCIAL," October 2019.
- [3] Cisco, "Cisco.com," 2022. [Online]. Available: <https://www.cisco.com/c/en/us/products/security/what-is-cybersecurity.html#~types-of-threats>.
- [4] S. & S. N. & M. S. P.S, "Overview of Cyber Security," IJARCCCE, vol. 7, pp. 125-128, 2018.
- [5] A. MAURI, "MyIDEAS," Departmental Working Papers, 04 2003. [Online]. Available: <https://ideas.repec.org/p/mil/wpdepa/2008-11.html> .
- [6] K. Adane, "SSRN," Arba Minch University, 1 September 2020. [Online]. Available: <https://ssrn.com/abstract=3545189>.

- [7] C. P. S. Technologies, "Linkedin," 26 November 2022. [Online]. Available: <https://www.linkedin.com/showcase/check-point-software-africa/?originalSubdomain=za>.
- [8] B. Gezahegn, "Assessment on Challenging Threats of Cyber Security and its Emerging trends on Selected Ethiopian Banking," 2020.
- [9] T. Shimels, "Maturity of Information System Security In Selected Private banks in Ethiopia," 2021.
- [10] K. Markos, "Factors influencing information security investments in the Ethiopian Banking sector," 2022.
- [11] B. Wedelu, "Information and Cyber Security Risk Assessment Framework for the Banking Sector in Ethiopia," 2017.
- [12] B. G. Mawudo, M.-H. Kim and B. G. Mawudor, "Continuous Monitoring Methods as a Mechanism for Detection and Mitigation of Growing Threats in Banking Security System," 2015.
- [13] A. Desisa, "Internet Banking Security Framework: The case of Ethiopian Banking Industry," 2014.
- [14] R. Chatterjee, "CIM," 5 MARCH 2020. [Online]. Available: <https://analyticsindiamag.com/difference-between-cybersecurity-information-security/>.
- [15] Kaspersky, "Kaspersky," 2022. [Online]. Available: <https://www.kaspersky.com/resource-center/definitions/what-is-cyber-security>.
- [16] NIST, "NIST COMPUTER SECURITY RESOURCE CENTER," 2013. [Online]. Available: <https://csrc.nist.gov/glossary/term/cybersecurity>.
- [17] A. O. N. B. A. Obotivere, "Cyber Security Threats on the Internet and Possible Solutions," International Journal of Advanced Research in Computer and Communication Engineering, vol. 9, 9, September 2020.
- [18] M. F. Hasan, "Cyber-attacks and Cyber Security Readiness: Iraqi Private Banks Case," Researchgate, August 2021.
- [19] M. Stamp, Information security:, 2011.
- [20] S. C. B. J. H. D. E. G. B. T. W. C. and C. M. , "Botnets: The Killer Web Applications," Elsevier, 2007.
- [21] M. H. "google," 2007. [Online]. Available: <http://books.google.com.my/books?id=MDzio>.
- [22] CISA, "Cyber security and Infrastructure security Agency," 20 November 2019. [Online]. Available: <https://www.cisa.gov/uscert/ncas/tips/ST04-015>.
- [23] Redhat, "www.redhat.com," 30 September 2022 . [Online]. Available: <https://www.redhat.com/en/topics/security/software-development-lifecycle-security>.
- [24] VMware, "Modern Bank Heists 4.0," 2021. [Online]. Available: <https://www.vmware.com/learn/security/modern-bank-heists-2021.html>.
- [25] J. M. B. a. F. T. Chong, "Emerging Technologies for Quantum Computing," vol. vol. 41, pp. pp. 41-47, Sept.-Oct. 2021, .

- [26] Techtarget, "Techtarget," [Online]. Available: <https://www.techtarget.com/searchcloudcomputing/definition/cloud-computing>.
- [27] S. Scorecard, "Security Scorecard," 07 12 2021. [Online]. Available: <https://securityscorecard.com/blog/6-strategies-for-cybersecurity-risk-mitigation/>.
- [28] E. Taddele, "ERP Post-Implementation Management Framework: The case of," Addis Ababa University, June 2015.
- [29] M. Velychko, "Conceptual Framework in Research," 2019.
- [30] Imperva, "Imperva," 2021. [Online]. Available: <https://www.imperva.com/blog/2021-in-review-part-1-5-cybersecurity-topics-that-made-news/>.
- [31] K. Noor, "Case Study: A Strategic Research Methodology: Universiti Industri Selangor," Shah Alam, pp. 1602-1904, 2008.
- [32] I. Akhtar, "Research Design," Social research foundation, September 2016.
- [33] T. Boru, "Mitigating cyber threats to banking industry," Universiti Utara Malaysia, 2018.
- [34] S. McCombes, "Correlational Research," [Online]. Available: <https://www.scribbr.com/methodology/correlational-research>. [Accessed 2019].
- [35] M. Given, "The SAGE Encyclopedia of Qualitative Research Methods," SAGE Publications, 2008.
- [36] Vedantu, "Vedantu," 2022. [Online]. Available: <https://www.vedantu.com/commerce/census-and-sample-survey>.
- [37] Simplilearn, "Simplilearn," 22 November 2022. [Online]. Available: <https://www.simplilearn.com/what-is-data-collection-article>.

Predicting Individuals' Vulnerability to Social Engineering Attacks

Tesfaye Abaya

HiLCoE School of Computer Science & Technology

jadon.tesfaye@gmail.com

Tibebe Beshah

School of Information Science, Addis Ababa University

tibebe.beshah@aau.edu.et

Abstract

The increasing prevalence of social engineering attacks (SEAs) poses a significant threat to individuals and organizations [1]. This study aims to develop and evaluate a predictive model for identifying individuals susceptible to SEAs. A dataset comprising 505 responses from university and colleges students in Addis Ababa, Ethiopia, was collected through an online survey. To address data insufficiency, synthetic data generation was employed, expanding the dataset to 3000 instances [2]. Employing a Design Science Research (DSR) methodology, multiple machine learning algorithms were compared, with Random Forest demonstrating superior performance in predicting SEA vulnerability. Key predictors identified include age, income, occupation, curiosity, and trust factors. The findings contribute to enhancing cybersecurity measures by enabling targeted interventions to protect vulnerable individuals.

Keywords- Social engineering attacks, Predictive model, Machine Learning, Random Forest, Cybersecurity

1. Introduction

The escalating reliance on digital technologies has ushered in an era characterized by unprecedented connectivity and data exchange. Concurrently, the landscape of cyber threats has evolved, with social engineering attacks (SEAs) emerging as a formidable challenge [3] [4]. These attacks, which exploit human psychology rather than technical vulnerabilities, have proven highly effective in compromising individuals and organizations alike. Despite advancements in cybersecurity, the human element remains a critical weak link [4].

The increasing sophistication of SEAs, coupled with the lack of comprehensive research on individual vulnerability, underscores the urgency of this study. Existing research often overlooks the complex interplay between demographic, psychological, and behavioral factors that contribute to susceptibility. This research aims to develop and evaluate a predictive model capable of identifying individuals at higher risk of falling victim to SEAs. By understanding these vulnerabilities, targeted interventions can be implemented to enhance cybersecurity defences.

To achieve this objective, a dataset comprising 505 responses from university students in Addis Ababa, Ethiopia, was collected through an online survey. Due to data insufficiency, synthetic data generation techniques were employed to expand the dataset to 3000 instances [2]. A Design Science Research (DSR) methodology was adopted for model development and evaluation. Multiple machine learning algorithms were compared, with Random Forest demonstrating superior performance in predicting SEA vulnerability.

2. Literature Review

The escalating prevalence of social engineering attacks (SEAs) has prompted extensive research into identifying factors influencing individual vulnerability. Previous studies have explored various dimensions of user susceptibility, including demographic, psychological, and behavioral factors.

Albladi and Weir [5] investigated the impact of user behavior, perceptions, and socio-emotional factors on social engineering vulnerability within social networks. Their study revealed that factors such as social media involvement, risk perception, and trust influence susceptibility. However, the focus on a specific social media platform and academic population limits the generalizability of their findings.

Huseynov and Ozdenizci [6] employed machine learning algorithms to predict social engineering susceptibility based on demographics, technology usage, and personality traits. While their study demonstrated the potential of machine learning in this domain, it focused primarily on email and SMS phishing, limiting its scope.

Wang et al. [7] examined the role of demographics, personality, cognitive processes, knowledge, and experience in predicting phishing vulnerability. Their research highlighted the importance of these factors in susceptibility but was constrained by a focus on email phishing. These studies collectively contribute to our understanding of social engineering vulnerability. However, they also reveal significant gaps in knowledge. Existing research often adopts a narrow perspective, focusing on specific attack types or limited demographic groups. Furthermore, the role of app usage behavior and social media habits in influencing susceptibility remains underexplored.

To address these limitations, this study expands the scope of inquiry by investigating a broader range of user characteristics. By incorporating app usage behavior, social media habits, and diverse demographic factors, this research aims to provide a more comprehensive understanding of individual vulnerability to SEAs. Additionally, the study employs a comparative analysis of multiple machine learning algorithms to identify the most effective model for predicting susceptibility.

This research builds upon the foundation laid by previous studies while venturing into new territory to provide a more comprehensive picture of social engineering vulnerability.

3. Methodology

3.1 Overview

This section outlines the methodology employed to develop a predictive model for assessing individual vulnerability to social engineering attacks (SEAs). It encompasses data collection strategies, sampling methods, experimentation techniques, and evaluation metrics utilized to assess the model's performance.

3.2. General Approach and Research Design

This research adopts a Design Science Research (DSR) framework, specifically following the process outlined by Offermann et al [8]. DSR is characterized by its iterative nature, focusing on the creation and evaluation of innovative artifacts to resolve real-world challenges. The DSR process consists of several key phases: problem identification, design and development, experimentation, and evaluation [8].

3.2.1 Problem Identification

The research commenced with an extensive literature review to identify existing challenges in cybersecurity, particularly the rising incidence of social engineering attacks, which are reported to constitute up to 98% of cyber incidents globally [9]. Insights from interviews with cybersecurity professionals further underscored the urgency of addressing the vulnerabilities associated with these attacks, particularly in the Ethiopian context, where tactics are becoming increasingly sophisticated.

3.2.2 Solution Design

The solution design phase involved several critical steps:

Data Preparation and Exploration

The core artifact of this research is a predictive model aimed at identifying individuals at risk of SEAs. The development process includes:

Technology Stack: The implementation utilized Python, leveraging libraries such as scikit-learn for machine learning, alongside tools like Visual Studio and Jupyter Notebook for data exploration and coding.

Data Acquisition and Synthetic Data Generation: Due to the lack of publicly available datasets relevant to the Ethiopian context, an online survey was conducted, yielding 505 responses. To enhance this dataset, the SMOTE (Synthetic Minority Over-sampling Technique) algorithm was applied, increasing the dataset to 3,000 instances. This step was crucial for balancing the dataset and improving model accuracy [2].

Data Preprocessing: The collected data underwent preprocessing to prepare it for machine learning applications. The dataset was divided into training (80%) and testing (20%) subsets.

Model Development

This stage focused on training and testing the predictive model using three machine

learning algorithms: Logistic Regression (LR), Decision Tree (DT), and Random Forest (RF). Each algorithm was rigorously tuned through hyperparameter [10] optimization to enhance performance. This comprehensive approach aimed to provide insights into the model's capabilities across various scenarios.

3.3 Evaluation

The evaluation of the model's performance involved several established metrics, including accuracy, precision, recall, and F1 score, particularly suited for binary classification tasks. To ensure robust performance estimates and mitigate overfitting, the following techniques were employed:

Data Splitting: The dataset was divided into training and testing sets using an 80/20 split, ensuring that the model was trained on a representative sample while retaining a separate set for unbiased evaluation.

Cross-Validation: A 5-fold cross-validation method was implemented, allowing for iterative training and evaluation on different data subsets. This technique enhances the model's generalizability to unseen data.

After conducting the evaluations, the findings were meticulously documented and analyzed, providing a comprehensive understanding of the model's strengths, weaknesses, and potential limitations. The F1 macro score was selected as the primary evaluation metric, reflecting the model's performance in predicting individual vulnerability to social engineering attacks.

4. Solution Design/ Data Experimentation/- Model Development

4.1 Experimental Setup

Data Splitting:

To ensure a fair evaluation of our machine learning models, we divided our dataset of **3000 instances** into training and testing sets. The **training set** comprised **80%** of the data, which was used to train the models. The remaining **20%** was allocated to the **testing set**, serving as a holdout set for evaluating the models' performance on unseen data.

Class Distribution Analysis:

Before proceeding with the experimentation, we carefully analyzed the class distribution within the training and testing sets. The specific class labels "high," "moderate," and "low" were defined based on domain expertise and predefined criteria, such as the frequency of phishing attempts, password breaches, or social media scams.

Table 1 Training and Test Data Class Distribution

Class	Training Data (Instances)	Training Data (Percentage)	Testing Data (Instances)	Testing Data (Percentage)
Low	789	32.90%	199	33.20%
Medium	1004	41.80%	248	41.30%
High	607	25.30%	153	25.50%

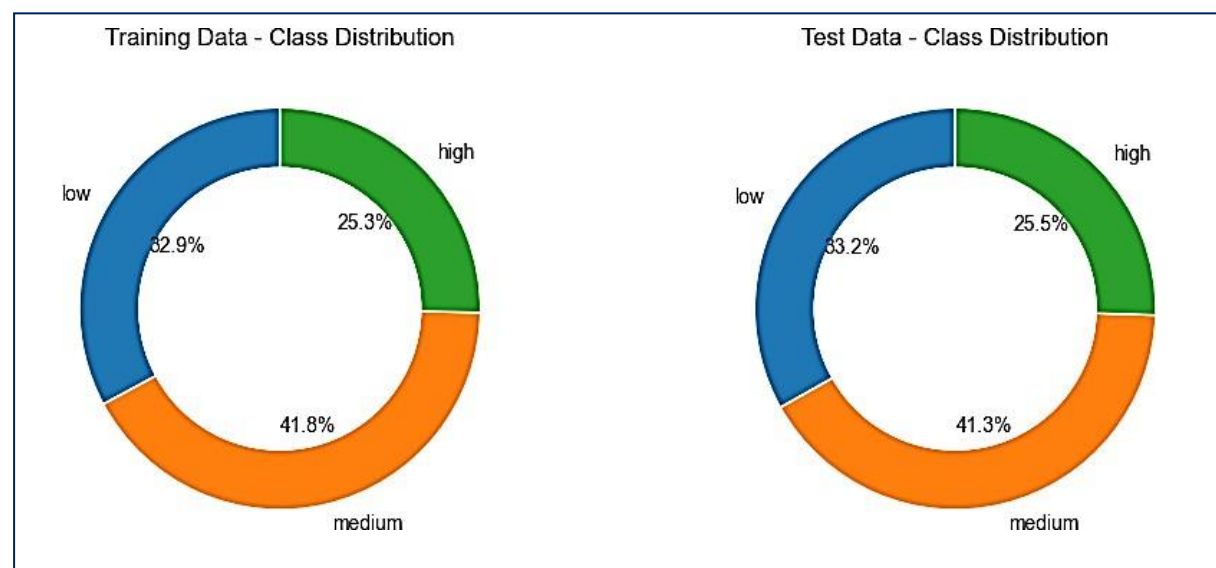


Figure 1 Training and Test Data Class Distribution

Evaluation Metrics:

We selected the following evaluation metrics to assess the performance of our machine learning models:

- **Accuracy:** This measures the overall proportion of correct predictions.
- **Precision:** This indicates the proportion of positive predictions that were actually correct.
- **Recall:** This measures the proportion of actual positive cases that were correctly predicted.
- **F1-score:** This is the harmonic mean of precision and recall, providing a balanced measure of both.
- **Cross-Validation:** To obtain reliable performance estimates and mitigate biases, we incorporated cross-validation techniques. In this study, we utilized a 5-fold cross-validation scheme, where the data was split into five subsets and the model was trained and evaluated on each subset in turn. This approach ensured a more

robust evaluation of the model's generalizability.

- **Hyperparameter Tuning for Model Optimization**

Hyperparameter [10] tuning was a critical component of our research. By carefully adjusting the hyperparameters of logistic regression, decision trees, and random forests, we were able to optimize their performance for social engineering vulnerability prediction. This process ensured a balance between model complexity and generalizability, allowing us to extract the maximum potential from each model.

Through hyperparameter tuning [10], we were able to tailor each model's configuration to the specific characteristics of our dataset. This customization helped us identify the most effective model for our task, ensuring that our final choice was not solely based on default settings. By mitigating the influence of hyperparameter choices on the final performance comparison, we were able to make a more objective and fair selection of the best model for social engineering vulnerability prediction.

4.2 Laboratory Experiment

4.2.1 Model 1 Logistic Regression

This research leverages logistic regression (LR) to develop a predictive model for assessing individual vulnerability to social engineering attacks (SEA). To address the challenge of missing data, we employed mean imputation to replace missing values with the average of observed values within each feature. This enriched dataset allowed the LR model to capture a broader range of relationships and interactions between variables, potentially improving its predictive accuracy and generalizability.

Data Splitting and Model Training

We utilized a dataset of 3000 instances, split into an 80/20 ratio for training and testing. This division provided sufficient data for model training while reserving a significant portion for unbiased evaluation.

Model Performance:

Our LR model, optimized through hyperparameter tuning, achieved a moderate level of predictive capability. The model exhibited an accuracy of 75.17%, correctly classifying approximately three-quarters of the data points. However, a more nuanced analysis revealed class-specific performance variations.

- **Class-Specific Performance:**

- **Class 0 (low vulnerable):** Achieved the highest precision (91.08%) and recall (94.76%), indicating a low rate of false positives and false negatives.
- **Class 1 (highly vulnerable):** Demonstrated the lowest precision (62.69%) and recall (55.63%), suggesting a higher risk of misclassification.

- **Class 2 (moderately vulnerable):** Achieved a moderate F1-score of 64.20%.

Impact of Class Imbalance:

The class-wise support values (number of instances) revealed an imbalanced distribution, with Class 1 having the highest support. This imbalance may have influenced the overall model performance, particularly for Class 2.

Our logistic regression model, optimized through hyperparameter tuning after data imputation, exhibits moderate predictive capability for identifying individuals vulnerable to social engineering attacks (SEA). While the achieved accuracy is encouraging, further investigation and potential refinements are needed to improve the model's generalizability and performance across all vulnerability categories.

Key Findings of LR:

- The LR model achieved an overall accuracy of 75.17%.
- Class-specific performance varied, with stronger results for low vulnerable individuals and weaker results for highly vulnerable individuals.
- Class imbalance may have influenced the model's performance, particularly for the highly vulnerable class.

4.2.2 Model 2 Decision Tree

In this study, after addressing missing values through data imputation, the second machine learning model we used to experiment on our data is decision tree. We applied hyperparameter tuning to a decision tree (DT) model to predict individual vulnerability to social engineering attacks (IVSEA). The tuning process focused on optimizing the model's performance by identifying the best values for parameters such as maximum depth, minimum samples per leaf, and minimum samples per split. The optimal hyperparameters were determined to be: `max_depth = 7`, `min_samples_leaf = 1`, and `min_samples_split = 2`.

The dataset, consisting of 3,000 instances, was divided into an 80/20 split for training and testing. This split ensured a robust training process while reserving enough data for unbiased model evaluation. The DT model achieved an accuracy of 86.17%, demonstrating strong predictive capability in identifying individuals vulnerable to social engineering attacks.

The model's performance across different classes varied. Class 1, which represented individuals most vulnerable to attacks, exhibited the highest precision (96.76%) and recall (96.37%), indicating accurate and consistent identification of this group. Class 2, representing a lower vulnerability, had the lowest precision (77.55%) and recall (75.50%), suggesting a higher rate of misclassification in this category. The F1-scores, which balance precision and recall, were highest for Class 1 (96.57%), followed by Class 0 (80.99%) and Class 2 (76.51%). Support values, representing the number of instances in each class, showed that Class 1 had the highest representation (248

instances), followed by Class 0 (199 instances) and Class 2 (151 instances). These values reflect the distribution of the dataset across different vulnerability categories.

In conclusion, the decision tree model, with optimized hyperparameters, showed strong overall performance in predicting individual vulnerability to social engineering attacks. However, the variation in precision, recall, and F1-scores across the classes indicates that further refinement of the model could improve its consistency, particularly for Class 2.

4.2.3 Model 3 Random Forest

In this study, a Random Forest (RF) model was implemented to predict individual vulnerability to social engineering attacks (IVSEA), with a particular focus on enhancing model performance through hyperparameter tuning. The dataset, comprising 3,000 instances, was divided into an 80/20 split for training and testing, ensuring a sufficient portion of data for model training while preserving a significant amount for unbiased evaluation.

Hyperparameter tuning played a crucial role in optimizing the Random Forest model. By adjusting key parameters such as the number of trees, maximum depth, and minimum samples per leaf, the model's performance was significantly enhanced, leading to superior predictive accuracy. The final model achieved an impressive accuracy of 99.67%, correctly classifying vulnerability status for nearly all instances in the dataset.

The precision and recall scores across all vulnerability classes were equally remarkable. For Class 0 (non-vulnerable individuals), both precision and recall were near perfect at 0.97, signifying that the model successfully identified almost all non-vulnerable cases. In Class 1 (vulnerable individuals), the model achieved a precision score of 0.992 and a perfect recall score of 1.00, indicating that it effectively captured all vulnerable individuals. For Class 2, both precision and recall scores were flawless at 1.00, highlighting the model's exceptional performance in this category.

The F1-scores, which measure the balance between precision and recall, were similarly outstanding. Class 0 achieved a perfect F1-score of 1.00, reflecting its strong classification performance. Class 1 had an F1-score of 0.996, and Class 2 achieved an F1-score of 0.993, further underscoring the model's balanced and robust predictive capability across all classes. Support values, representing the number of instances per class, revealed that Class 1 had the largest representation with 248 instances, followed by Class 0 with 199 instances, and Class 2 with 151 instances. This distribution helped illustrate the consistency of the model across varying class sizes.

Overall, the implementation of hyperparameter tuning significantly enhanced the Random Forest model's predictive power. The near-perfect accuracy, precision, recall, and F1-scores demonstrate the model's effectiveness in identifying vulnerable individuals, making it a reliable tool for predicting individual vulnerability to social engineering attacks. These findings suggest that the Random Forest model, when fine-

tuned with optimized hyperparameters, is an exceptionally accurate and robust method for addressing cybersecurity vulnerability prediction.

5. Results and Discussions

5.1 Results

This study aimed to assess the effectiveness of a Random Forest model with optimized hyperparameters in predicting individual susceptibility to social engineering attacks (SEAs). By fine-tuning the model's parameters, we sought to enhance its predictive power and identify key factors influencing vulnerability. The analysis revealed several significant predictors, ranked by their importance in the model:

Age: Age emerged as the most influential factor, contributing 38% to the model's predictive power. The findings indicate that older individuals are more susceptible to SEAs, highlighting the need for targeted interventions for this demographic.

Average Monthly Income: With a 10% importance value, income was another major predictor of vulnerability. The data suggests that individuals with higher incomes may be more prone to social engineering tactics, perhaps due to increased financial targets or overconfidence in their digital interactions.

Occupation: Occupation accounted for 6% of the predictive power, with individuals working in science and business fields being more vulnerable to SEAs. Further investigation is required to explore the specific factors within these professions that may contribute to this heightened susceptibility.

Curiosity: This trait was associated with a 5% predictive value. Individuals with higher levels of curiosity exhibited an increased susceptibility to SEAs, suggesting that attackers may exploit curiosity-driven behavior through deceptive tactics.

Trust in Apps: Trusting apps had a 5% importance value, indicating that individuals who have high trust in apps are moderately more vulnerable to SEAs. This underscores the need for awareness regarding app-related threats, as attackers may leverage users' trust in seemingly legitimate applications.

Trust in Friends from Contact Lists: Trust placed in contacts also had a 5% predictive value. This finding highlights the potential for social engineering attackers to exploit pre-existing trust relationships, often relying on contacts or impersonation tactics.

Self-Efficacy and Extroversion: Both traits contributed 4% to the model's predictive power. The data suggests that individuals with higher self-efficacy and extroversion may be more vulnerable to SEAs, possibly due to overconfidence or social interactions that create opportunities for manipulation.

Field of Study: Educational background, reflected by the field of study, also had a 4% predictive value. This suggests that certain educational disciplines may correlate

with varying levels of susceptibility to SEAs.

Duration of Social Media Usage: Daily social media usage contributed 3% to the predictive power. Individuals with higher usage showed a moderate increase in vulnerability, possibly due to increased exposure to attack vectors commonly found on social media platforms.

Frequency of App Updates: Surprisingly, the frequency of updating apps had a lower importance value (2%) in predicting SEA susceptibility. The data indicated that those who occasionally update their apps might be more vulnerable than those who update regularly, although the overall effect was minimal.

In summary, the optimized Random Forest model identified age (38%) and average monthly income (10%) as the most critical factors in predicting susceptibility to SEAs. These findings suggest that cybersecurity awareness initiatives could be particularly effective if tailored to these demographics, emphasizing the importance of educating older and higher-income individuals about the risks of social engineering. Additionally, while other factors such as occupation, curiosity, and social media usage play smaller roles (2% to 6%), they remain relevant in understanding overall vulnerability. The model's results provide a comprehensive framework for developing targeted interventions aimed at mitigating the risks posed by social engineering attacks.

5.2 Discussions

This study provides a comprehensive analysis of individual susceptibility to social engineering attacks (SEAs) using hyperparameter-tuned machine learning models (MLMs), particularly the Decision Tree (DT), Random Forest (RF), and Logistic Regression (LR) models. The focus was on identifying demographic, behavioral, and psychological factors that contribute to vulnerability, and evaluating the models' performance in predicting individuals' susceptibility to social engineering attacks in Ethiopia.

Key features, such as age, gender, educational level, field of study, occupation, average monthly income, social media usage patterns, and frequency of updating apps and operating systems (OS), were analyzed. These features exhibited similar distributions across all target classes, with some variations across different levels of vulnerability. The analysis found that factors like age, number of social media contacts, and duration of social media usage per day displayed distinct distributions across all target classes, indicating their significance in differentiating vulnerability levels.

In terms of model performance, the Random Forest (RF) model emerged as the best performer, achieving an accuracy of 99.67%, surpassing both the Logistic Regression (LR) model with 75.1% accuracy and the Decision Tree (DT) model with 86.7% accuracy. This high accuracy suggests that the RF model is highly effective in identifying individuals who are vulnerable to SEAs. The RF model's success can be attributed to hyperparameter tuning and data imputation, which improved its performance over other models.

Several factors emerged as key predictors of vulnerability. Age (38% importance value) and average monthly income (10%) were the most significant contributors. Older individuals and those with higher incomes were more vulnerable to SEAs, likely due to factors such as increased financial exposure and less familiarity with modern online threats. Occupation (6%), particularly individuals in science and business fields, also played a crucial role in determining vulnerability, likely due to the nature of their work making them more susceptible to certain social engineering tactics. Other factors such as curiosity (5%), trust in apps (5%), and trust in social connections (5%) demonstrated moderate predictive power, emphasizing the role of behavioral traits in influencing susceptibility.

The findings underscore the need for cybersecurity awareness programs tailored to these key factors, particularly focusing on older individuals and those with higher incomes, as well as individuals in specific professional fields. By addressing the behavioral tendencies that make individuals more vulnerable, such as curiosity and trust in apps and contacts, targeted interventions can be designed to improve resilience against SEAs.

While accuracy is a critical metric, other evaluation measures such as precision, recall, and F1- score should be considered to gain a comprehensive understanding of the model's performance. Future research should validate the findings across broader datasets and populations to ensure the model's generalizability and reliability in predicting individual vulnerability to SEAs in different contexts.

In conclusion, this research advances the understanding of social engineering attack vulnerability by identifying key demographic and behavioral factors that influence susceptibility. The Random Forest model, with its optimized hyperparameters, demonstrated superior performance in predicting vulnerability, highlighting its potential as a reliable tool for cybersecurity risk assessment. The findings also emphasize the importance of tailored security awareness programs that consider individuals' unique demographic and behavioral profiles to mitigate the risks posed by SEAs effectively.

6. Conclusions

The research successfully addressed the core questions and objectives outlined at the beginning of the study, providing valuable insights into the prediction of individual vulnerability to social engineering attacks (SEAs) using machine learning models (MLMs).

What type of individuals are most vulnerable to SEA?

This study identified several key demographic and behavioral characteristics that contribute to an individual's susceptibility to SEAs. Age emerged as the most influential factor, with older individuals being more vulnerable. Income, occupation, and curiosity also played significant roles, highlighting that individuals with higher incomes and those working in fields such as science and business were more likely to

fall victim to these attacks. Additionally, trust in apps and social connections further heightened vulnerability, as attackers exploit these trusted relationships.

Which factors primarily influence a person's vulnerability to SEA?

The research demonstrated that demographic factors like age and income are primary predictors of vulnerability, but behavioral traits such as curiosity, trust in technology, and social media usage patterns also significantly influence susceptibility. The importance of trust in apps and frequency of social media usage underlines the role of user behaviors in creating opportunities for social engineers to exploit.

Which MLMs are suitable for predicting social engineering attack vulnerability, and under what circumstances are they best suited for doing so?

In answering this question, the research compared multiple machine learning models, including Random Forest (RF), Logistic Regression (LR), and Decision Trees (DT). The Random Forest model, with hyperparameter tuning, emerged as the most effective model, achieving an accuracy of 99.67%, far surpassing the other models. This demonstrates that RF is highly suitable for predicting vulnerability to SEAs, especially in contexts where diverse demographic and behavioral data are available. The model's performance was further enhanced through the use of data imputation techniques, ensuring reliable predictions even in the presence of missing data.

Research Objectives Achieved:

The general objective of the study was to develop a model for predicting individual vulnerability to social engineering attacks and propose the most suitable machine learning model. This objective was fully achieved through the successful experimentation with various MLMs, identifying Random Forest as the best performer. The research further advanced the understanding of which demographic, behavioral, and personality traits are critical in determining SEA vulnerability, fulfilling the first two specific objectives.

The predictive model developed as part of the study is capable of identifying individuals who are most susceptible to SEAs based on factors such as age, income, curiosity, and trust in apps. This directly addressed the third specific objective. The final objective, examining which MLMs are most suitable, was thoroughly explored, with RF being identified as the optimal choice under the tested conditions.

In conclusion, this research provides a robust model for predicting individual vulnerability to social engineering attacks, with Random Forest proving to be the most suitable machine learning model for this task. The study successfully answered the research questions by identifying key factors influencing vulnerability and establishing a predictive framework that can be used for targeted security awareness and interventions. As a result, the research offers a meaningful contribution to both the academic understanding of social engineering vulnerabilities and practical efforts to mitigate them.

7. Future Work

This research has successfully identified key factors that contribute to individual vulnerability to social engineering attacks (SEAs), using machine learning models to analyze demographics, personality traits, and security awareness. Notably, age and income were found to be the most influential predictors, providing a foundation for developing targeted prevention strategies to mitigate these risks.

For future work, several opportunities arise to enhance the model and its practical applications. First, creating a user-friendly interface would enable organizations to easily assess the vulnerability of employees or customers, facilitating more tailored security awareness training programs. Second, refining the model by incorporating additional data sources, exploring new machine learning algorithms, and testing its generalizability across different populations will help improve its accuracy and versatility. Third, developing a dynamic risk assessment system that monitors real-time online behavior could allow for continuous vulnerability scoring, triggering timely interventions when an individual's risk increases.

Lastly, an exciting direction for future research involves expanding the model's capabilities to detect and respond to ongoing social engineering attacks in real-time. This would enhance the system's defensive capabilities, enabling organizations to not only predict vulnerabilities but also actively defend against evolving SEAs. Pursuing these avenues will further advance this research into a powerful, actionable tool that can be used by both individuals and organizations to proactively combat the growing threat of social engineering attacks.

References

- [1] S. Morgan, "Cybercrime To Cost The World 8 Trillion Annually In 2023," *Cybercrime Magazine*, Oct. 13, 2022. <https://cybersecurityventures.com/cybercrime-to-cost-the-world-8-trillion-annually-in-2023/>
- [2] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," *Journal of Artificial Intelligence Research*, vol. 16, no. 16, pp. 321–357, Jun. 2002, doi: <https://doi.org/10.1613/jair.953>.
- [3] Wikipedia Contributors, "Cyberattack," *Wikipedia*, Sep. 24, 2019. <https://en.wikipedia.org/wiki/Cyberattack>
- [4] JPMorgan, "How Criminals Use Social Engineering to Target Your Company."
- [5] S. Albladi and G. Weir, "Predicting individuals' vulnerability to social engineering attacks," *International Journal of Advanced Computer Science and Applications*, vol. 10, no. 2, pp. 144-152, 2019. [Online]. Available: [https://thesai.org/Downloads/Volume10No2/Paper_19-](https://thesai.org/Downloads/Volume10No2/Paper_19-Predicting_Individuals_Vulnerability_to_Social_Engineering_Attacks.pdf)
- [6] [Predicting_Individuals_Vulnerability_to_Social_Engineering_Attacks.pdf](https://thesai.org/Downloads/Volume10No2/Paper_19-Predicting_Individuals_Vulnerability_to_Social_Engineering_Attacks.pdf)
- [7] F. Huseynov and B. Ozdenizci Kose, "Using machine learning algorithms to

predict individuals' tendency to be victim of social engineering attacks," *Information Development*, p. 026666692211163, Aug. 2022, doi: <https://doi.org/10.1177/02666669221116336>.

[8] R. Yang, K. Zheng, B. Wu, D. Li, Z. Wang, and X. Wang, "Predicting User Susceptibility to Phishing Based on Multidimensional Features," *Computational Intelligence and Neuroscience*, vol. 2022, pp. 1–11, Jan. 2022, doi: <https://doi.org/10.1155/2022/7058972>.

[9] P. Offermann, O. Levina, M. Schönherr, and U. Bub, "Outline of a design science research process," *Proceedings of the 4th International Conference on Design Science Research in Information Systems and Technology - DESRIST '09*, 2009, doi: <https://doi.org/10.1145/1555619.1555629>.

[10] M. Security, "Are Social Engineering Attacks on the

[11] Rise?," www.mitnicksecurity.com.
<https://www.mitnicksecurity.com/blog/are-social-engineering-attacks-on-the-rise>

[12] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning*. New York, NY: Springer New York, 2013. doi: <https://doi.org/10.1007/978-1-4614-7138-7>.

Recommended Computer Aided Design Software for EASE Engineering PLC

Asmerom Estifanos
HiLCoE School of Computer Science & Technology
jadon.tesfaye@gmail.com

Abstract

The paper focuses on selecting the most suitable Computer-Aided Design (CAD) software for EASE Engineering PLC's post-tensioning design automation needs. AutoCAD, BricsCAD, ProgeCAD, and ZWCAD were evaluated based on their in-built functions, automation capabilities, and compatibility with EASE Engineering's drafting requirements. The findings consider affordability, API support, AutoLISP processing capacity, and overall suitability for the specific needs of post-tensioning design. It has been noted that ZWCAD is the most suitable option with regards to affordability and overall performance.

Keywords: CAD, AutoLISP, DCL, API, Dynamic block

1 Introduction

EASE Engineering PLC, is believed to be the first company that introduced design and construction of post-tensioning for building and bridge in Ethiopia. Being the pioneer in the field the company needs to implement all the efficient tools that are currently in the market in order to be ahead of competitors that are entering the market. A Computer Aided Design (CAD) software is among the different design tools the company uses. This CAD software is used to convert the design output of other structural analysis and design software into a structural drawing. This conversion is a very time consuming operation when done manually and often many of the tasks need to be automated with scripting languages that are shipped with most CAD tools.

The need for accuracy and speed in drafting work is growing in the Ethiopian construction sector, where EASE Engineering is a part of. The organization is always seeking for ways to optimize the drawing process for post-tensioning systems. Drafting tasks that are automated may result in fewer errors, more time saved, and better project outcomes overall. EASE Engineering is thus driven to investigate and contrast the in-built functions as well as automation capabilities of AutoCAD, BricsCAD, ProgeCAD and ZWCAD. Thus, the objective is to investigate and compare the in-built functions and automation capabilities of AutoCAD, BricsCAD, ProgeCAD, and ZWCAD.

2 Related Work

Computer-Aided Design (CAD) is a technology that revolutionized the process of designing and drafting by enabling engineers, architects, and designers to create precise digital representations of physical objects or structures. CAD software allows users to create, modify, analyze, and optimize designs with ease and accuracy. It has become an essential tool across various industries, including automotive, aerospace, architecture, construction, electronics, and manufacturing.

CAD software provides a virtual workspace where designers can sketch, model, and simulate their ideas in 2D or 3D formats. These tools offer a wide range of features, including parametric modeling, which allows users to define geometric relationships and constraints to create dynamic designs that can be easily modified. Additionally, CAD software often integrates with other technologies such as computer-aided manufacturing (CAM) and computer-aided engineering (CAE), facilitating seamless workflows from design to production.

CAD software vary in complexity and capability, catering to different needs and skill levels. Some popular CAD programs include SolidWorks, Inventor, and AutoCAD, each with its own strengths and specialties. These tools empower designers to visualize concepts, detect errors early in the design process, improve collaboration among team members, and ultimately bring innovative ideas to life with efficiency and precision. It is apparent that CAD tools have made production of design drawings very efficient as opposed to the manual production of design drawings. One can use the array of functions that are shipped with these tools. However, the environment of construction greatly varies and the functions in these tools are seldom comprehensive. Different authors try to address the further automation of these tools using APIs. The research by Sheng Jin et al. [1] proposes an innovative method for automatic CAD design in construction engineering. The method uses on-site geomorphological data to generate 3D CAD models and 2D drawings, reducing labor and time costs. Implemented through the API of CAD systems, the method also incorporates mixed reality technology to validate assembly results, providing a realistic and efficient experience for users and customer. Li et al. [2] try to examine for instance the C# .NET API of AutoCAD in creation of plugins and generation of graphics. This article shows that CAD software might have insufficient or missing functions required by some fields as compared to others. Therefore, it is necessary to carry out secondary development of CAD and design professional modules that conform to specific fields [2]. This can extend the application range of CAD, and can also improve the designer's work efficiency and reduce the errors in design, which has strong practical value.

A study by Anastasios et al. [3] examines the creation of an application that automates

the design process for a typical pneumatic double-acting cylinder based on predetermined parameters. The authors show that for repetitive tasks such as building the double-acting cylinder can be overwhelming for a drafter and takes too much time that could be better allocated. So using APIs one can draw the cylinder instantly by just using a dialog box to enter the required parameters such as diameter and with the click of a button the required graphics can be generated. Applications like these can aid the productiveness of many designers and labor forces that produce a limited range of products or that are specialized in a specific range of mechanical systems (i.e. machinery) [3]. Another paper by I. A. Magomedov et al. [4] tries two CAD tools that were developed by two different companies for the purpose of 2D/3D modeling and analyzing. They compared two software, Solidworks (by SolidWorks Corp.) and Inventor (by Autodesk Inc.). The parameters chosen as the basis of comparison in this paper were functionality, market share, user friendliness and availability. The authors underline the importance of correct outcome from the software packages in engineering when choosing software.

3 Comparative Analysis

Among the varieties of options provided by the providers, the following are used for the purpose of this comparison AutoCAD 2022, BricsCAD Ultimate 2024, ProgeCAD Professional 2024 and ZWCAD Professional 2024. All of the drafting work in EASE Engineering PLC is done in 2D, hence, our main focus in this paper will be in 2D aspects of the programs even though all four programs have 3D drafting capabilities. In addition Windows OS is used as the basis for comparison because it is the main OS used in the company.

According to Open Design Alliance (ODA), a professional application development platform, the source code they provide to members of their consortium consists of 20M+ Lines of code, 850 Person-years in development and 45 Projects. One of the ODA members, IntelliCAD Technology Consortium (ITC), provides a CAD programming platform as well. IntelliCAD is based on an API provided by ODA and expands on it. Developing a CAD software from the ground up is evidently a very expensive task. And as a result, BricsCAD uses the ODA technology while ProgeCAD and ZWCAD are powered by IntelliCAD. Both ODA and ITC are not affiliated with Autodesk Inc. who is the owner of AutoCAD.

The applications will be compared based on their capability in:

- Interface and Usability: User interface and the ease of getting used to the applications.
- Price and Licensing: Modalities of licensing and affordability.
- Features and Capabilities: In-built functions that are shipped with the products.
- Compatibility: File formats that are supported by these products.
- System Requirements: Resources required to run the applications.

- **API Support:** Programming languages supported to be able to introduce user defined functions and extend application capabilities.
- **Performance and Speed:** Execution times required by the applications for tests that are performed using a custom made (built with AutoLISP API) program.

The applications under investigation are evaluated against these parameters and the results obtained are presented below.

Interface and Usability: All three of the CAD applications have a user interface that is similar to the pioneer application, AutoCAD. They all come with command lines, toolbars and menus which are more or less in similar positions, with a customizable interface. Anyone who has been using AutoCAD, will not be facing a steep learning curve to switch to any of the three applications.

Price and Licensing: Pricing as well as modalities of payments greatly vary among these programs. The licensing is no exception, licensing ranges from subscription to perpetual.

Table 1: Pricing and licensing

Application	Perpetual	1 month	1 year	3 year	Flex
AutoCAD	-	\$245	\$1975	\$5925	Min - \$300 for 100 tokens (Pricing is variable)
ZWCAD	\$1199	-	\$432	\$1166	-
ProgeCAD	\$399	-	-	-	-
BricsCAD	\$2100	-	\$1120	\$3020	-

AutoCAD used to provide perpetual licenses, however, as of January 2016 it has stopped selling new perpetual licenses and moved to plan subscription instead. In addition it has introduced a new method which it calls, Flex. Flexes can be used as per usage that can be useful for users who want to use a specific tool for short periods at times, with 7 tokens/day. And on the other hand ProgeCAD only provides perpetual license and no additional renewable fees as can be seen in Table 1.

Compatibility: Design drawings used in the construction industry among many others use the unofficially standardized DWG file format. DWG is the native file format for AutoCAD, developed by Autodesk. All four applications work with the DWG file format. Import and export with one another work seamlessly. In addition, construction drawings are exchanged between multiple stakeholders of a project very frequently. AutoCAD is the most widely used application in Ethiopian construction industry and compatibility of other CAD

applications with AutoCAD carries a lot of weight in the selection of the most suitable software for EASE Engineering PLC.

System Requirements: AutoCAD has been around for over 40 years and has acquired a large user base and significant resources. As a result it has extensive feature set and a wide range of tools, which has caused it to demand more computational resource than any of the alternative CAD applications chosen for this study. As can be seen from Table 4.4.1 AutoCAD requires a minimum disk space of 10 GB as opposed to the 2 GB required by ZWCAD/BricsCAD and 4 GB by ProgeCAD. The memory requirement is no exception, it requires 4 times that of what is required by ZWCAD. However, most computers that are commercially available these days are not resource deficient. And consequently, a user is more concerned with the availability of tools and performance.

Table 2: System Requirements

Requirement	AutoCAD	ZWCAD	ProgeCAD	BricsCAD
Operating System	64-bit Microsoft® Windows® 11 and Windows 10	Microsoft® Windows 7 sp1 and above	Microsoft® Windows 7 and above	64-bit Microsoft® Windows® 11 and Windows 10
Processor	2.5–2.9 GHz processor	Intel® Pentium™ 4 1.5 GHz or equivalent AMD® processor	2.2 GHz or faster 64-bit (x64) processor	Intel® Core™ i5 AMD Ryzen™ 5
Memory	8 GB	2 GB	4 GB	8 GB
Display Resolution	1920 x 1080 with True Color	1024*768 VGA with True Color	-	Full HD (1920 x 1080) True Color display
Display Card	1 GB GPU with 29 GB/s Bandwidth and DirectX 11 compliant	1GB, support OpenGL 4.2	-	1 GB VRAM
Disk Space	10.0 GB	2 GB free space in OS disk, 1 GB free space in installation disk	4 GB	2 GB
.NET Framework	.NET Framework version 4.8 or later	-	-	-

Features and Capabilities: In general AutoCAD has by far the most features, hence, the higher disk space requirement shown in Table 4.4.1. However, for the case of this comparison which only focuses on the 2D features, all four programs achieve the basic requirements, except for some differences in the way they each handle dynamic blocks, a feature that's frequently used in the drafting of post tensioning strands as a way to quickly change between live and dead end representations. Dynamic blocks are a proprietary pieces of Autodesk. All of the other three programs do not allow directly editing of dynamic blocks made with AutoCAD. Nonetheless, ZWCAD has its own alternative which it calls flexblocks. Flexblocks function the same way the AutoCAD dynamic blocks do, however, in order to edit dynamic blocks created with AutoCAD it must first be converted to ZWCAD's native type as shown in Figure 1.

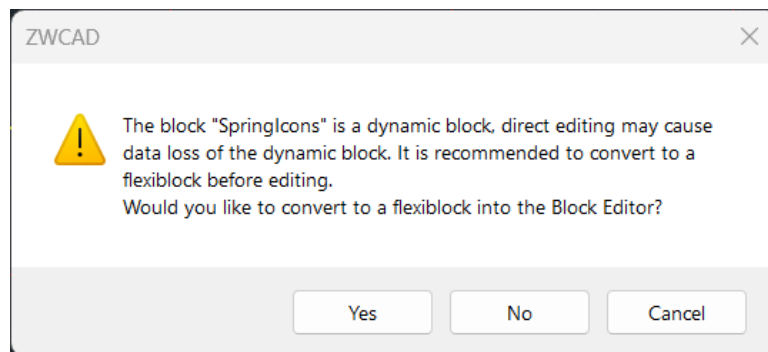


Figure 1: ZWCAD's prompt when attempting to edit a dynamic block

BricsCAD also allows to fully utilize dynamic blocks created with AutoCAD, however, in order to edit them they must first be converted to BricsCAD's native type called parametric blocks as shown in Figure 2. Both ZWCAD and BricsCAD perform the conversion automatically based on the user's preference.

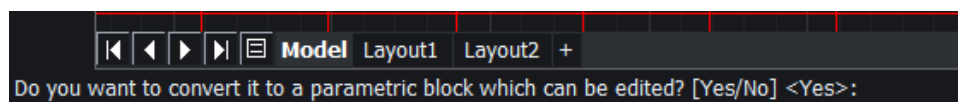


Figure 2: BricsCAD's prompt when attempting to edit a dynamic block

ProgeCAD has no alternative entities. Figure 3 shows the message that is generated by ProgeCAD to use the REFEDIT command when a user attempts to edit a dynamic block created by AutoCAD, which makes the dynamic block lose its dynamic property and become a regular block.

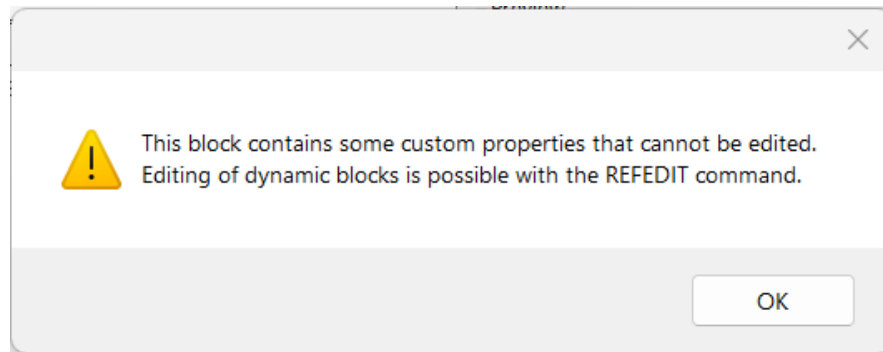


Figure 3: ProgeCAD's message when attempting to edit a dynamic block

Their capabilities can be compared with regards to their support to APIs which is regarded as the gateway to extend and test the capabilities of the programs with regards to the user's specific needs.

API Support: There are many APIs that are shipped with most of the CAD applications. However the most commonly used languages currently are C++, .NET, VBA, LISP and DCL. Table 3 shows the availability of those APIs in each applications.

Table 3: API Support

Language	AutoCAD	ZWCAD	ProgeCAD	BricsCAD
C++	✓	✓	-	✓
.NET	✓	✓	✓	✓
VBA	✓	✓	✓	✓
LISP	✓	✓	✓	✓
DCL	✓	✓	✓	✓

Two of the APIs, List Processing (LISP) and Dialog Control Language (DCL) have been chosen to create the logic and dialog boxes, respectively, in order to test the performance and speed of the programs.

Performance and Speed: Two of the most repetitive tasks that are done in EASE Engineering PLC are calculating tributary areas in order to calculate what portion of the slab loads are transferred to the corresponding column below and to consequently calculate check for punching shear. This task is frequently performed to quickly check software outputs and to have quick estimations manually. And the other repetitive task is the comparison of different versions of drawings to check for shifts in column locations and changes in slab layouts.

These operations have been automated by the author of this paper using AutoLISP. AutoCAD's IDE for AutoLISP was used for coding and debugging. A key issue that was noticed in this phase were:

- AutoCAD and BricsCAD both possess an internal IDE for AutoLISP.
- ZWCAD does not ship with its own AutoLISP IDE but can connect to Visual Studio and use it for coding and debugging
- Even though ProgeCAD can accept AutoLISP files and run them (to a certain extent) it was noticed that it does not ship with its own IDE. And no documentation has been found whether or not it can use external IDEs for coding and debugging.

A runtime analysis was performed by registering the start and completion times of the scripts on the machine the applications are being tested on in order to have an approximate execution time. Therefore, it should be noted that the duration mentioned in the paper is specific to the machine it was tested on. Regardless of the effort to keep the machine in similar state, it should be noted that the state of the machine is prone to changes as a result of factors that cannot be controlled and could have an effect on the measure of the time a task is completed.

Specification of the machine used for testing was:

- Processor - 11th Gen Intel(R) Core(TM) i7-11800H @ 2.30GHz
- RAM - 16.0 GB
- OS - Windows 11 Pro, 64-bit operating system, x64-based processor

Tributary areas

For this script an algorithm that is used to produce Voronoi Diagrams have been adopted. The script implements the drawing of lines, creating boundary polylines, inserting points, deleting entities and inserting blocks. Rectangular slab shown in Figure 4 consisting of 459 columns (blocks) is used for the comparison of performance.

Execution time recorded:

- AutoCAD – 22 seconds (Figure 5)
- ZWCAD – 2.22 minutes (Figure 6)
- BricsCAD – 49 seconds (Figure 7)
- ProgeCAD – The script generated stack overflow error and aborted (Figure 8)

Dialog Box Design

Dialog Control Language (DCL) is used to check the ability of each application to produce the expected dialog boxes. Figures 9 - Figure 12 show that all four applications

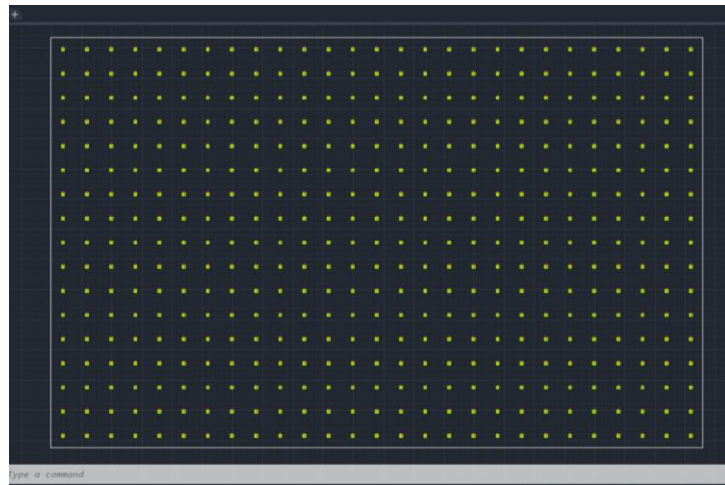


Figure 4: Slab and column layout before “Tributary area” script is run

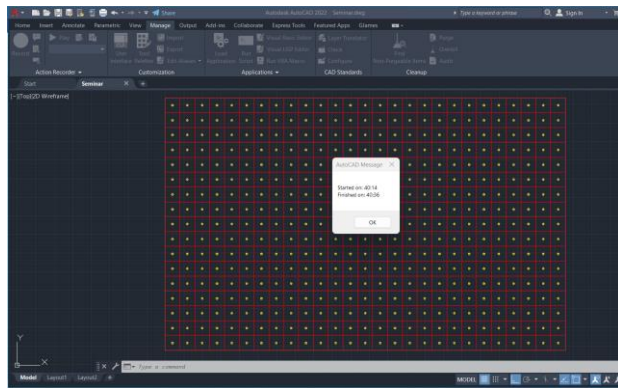


Figure 5: AutoCAD output after the “Tributary area” AutoLisp script is run

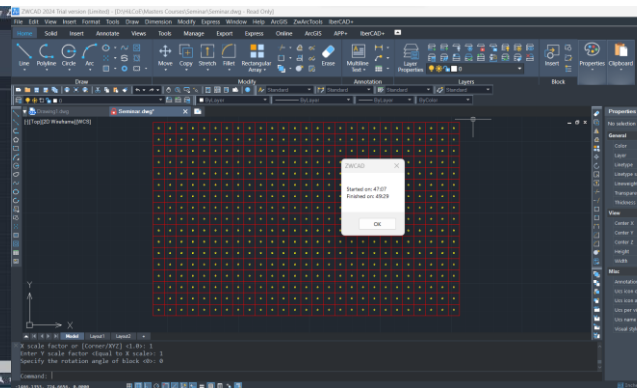


Figure 6: ZWCAD output after the “Tributary area” AutoLisp script is run

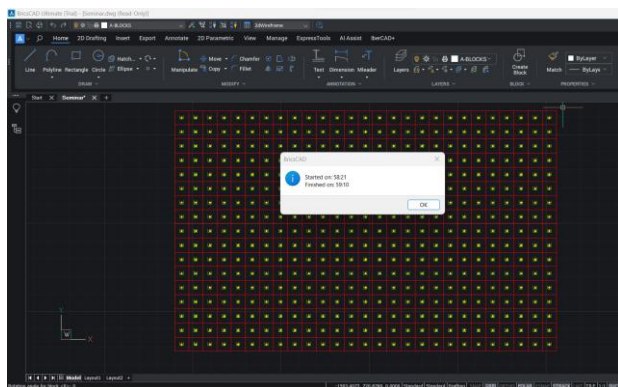


Figure 7: BricsCAD output after the “Tributary area” AutoLisp script is run

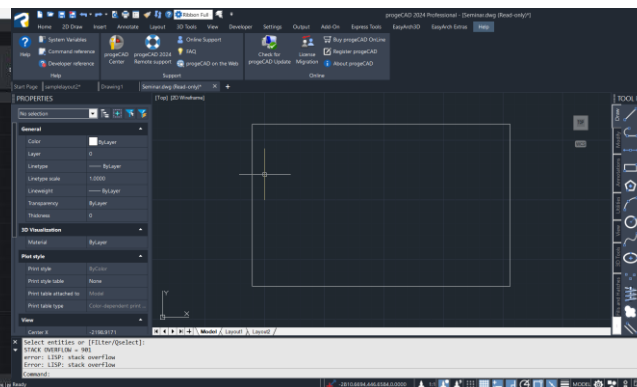


Figure 8: ProgeCAD output after the “Tributary area” AutoLisp script is run

displayed the dialog box as required with minor misalignment of components observed.

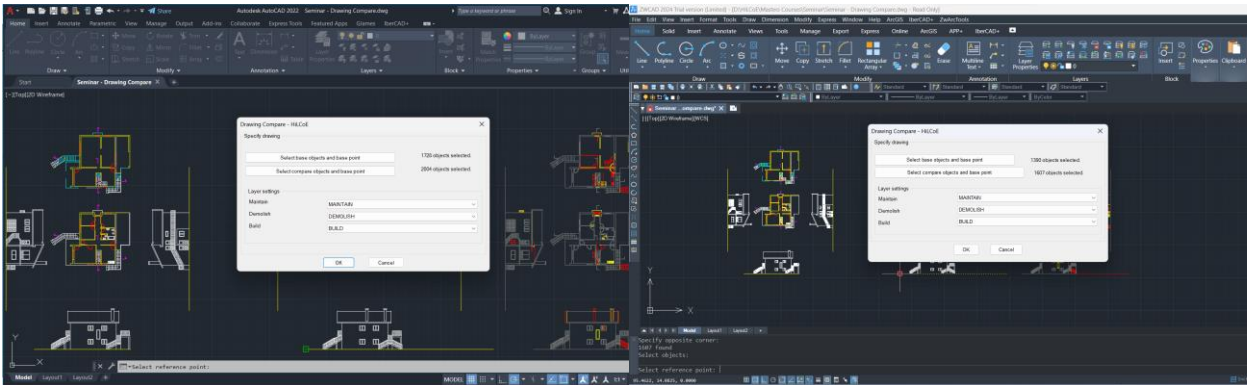


Figure 9: AutoCAD dialog designed with DCL

Figure 10: ZWCAD dialog designed with DCL

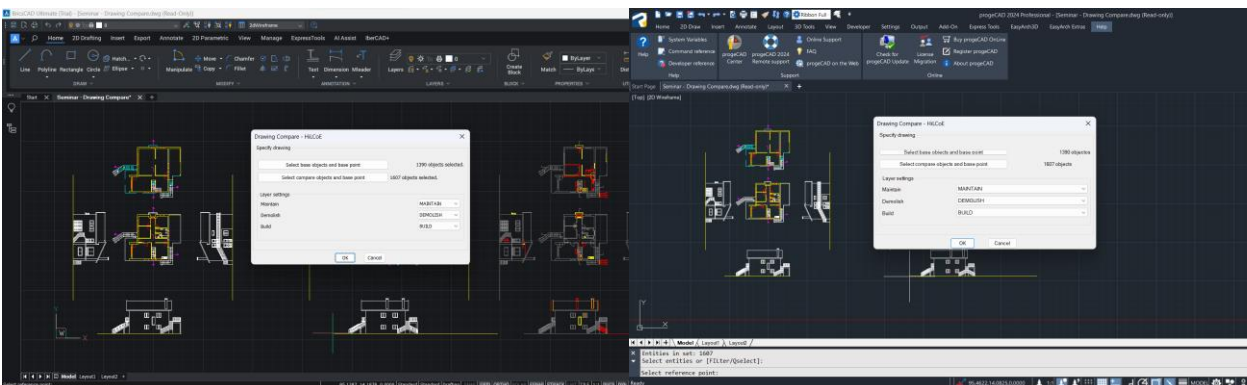


Figure 11: BricsCAD dialog designed with DCL

Figure 12: ProgeCAD dialog designed with DCL

The final script used for the measurement of the performance is a script that compares two drawings and produce a third drawing that has a color code of Gray (unchanged), Yellow (Exists in the first drawing, base objects, but not in the second) and Red (exists in the second drawing, compared objects, but not in the first). The two drawings to be compared are shown in Figure 4.7.3.1, base base objects are in the left (1390 objects) and compared objects (1607 objects) are the ones on the right. And the drawing produced by the “Drawing Compare” script can be seen in the far right side of Figures 13 – Figure 17. Three of the applications produced the expected output while ProgeCAD aborted the routine without producing a result as shown in Figure 17.

Execution time recorded:

- AutoCAD – 31 seconds (Figure 14)
- ZWCAD – 4 seconds (Figure 15)
- BricsCAD – 1 second (Figure 16)

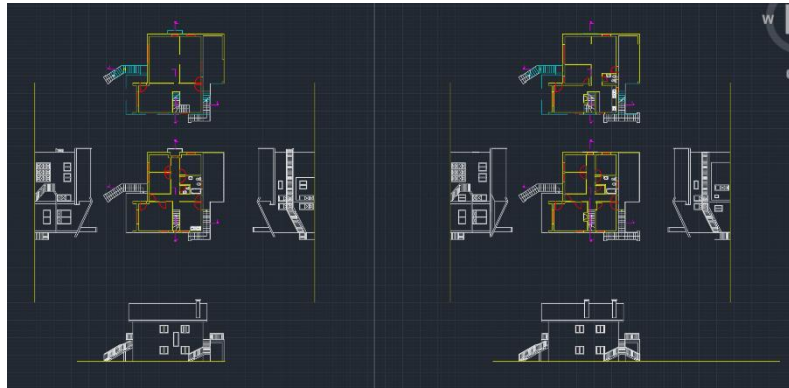


Figure 13: Drawings to be compared by AutoLISP “Drawing Compare” script

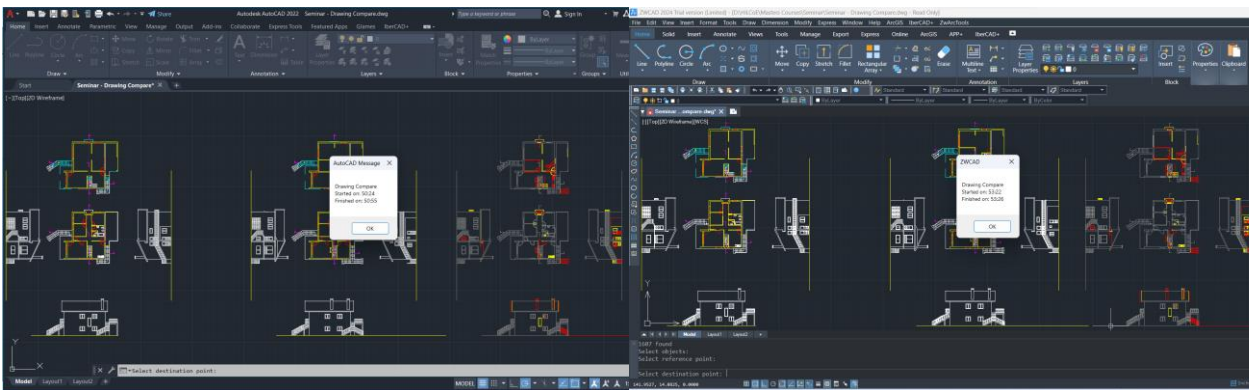


Figure 14: AutoCAD output after “Drawing Compare” AutoLISP script is run

Figure 15: ZWCAD output after “Drawing Compare” AutoLISP script is run

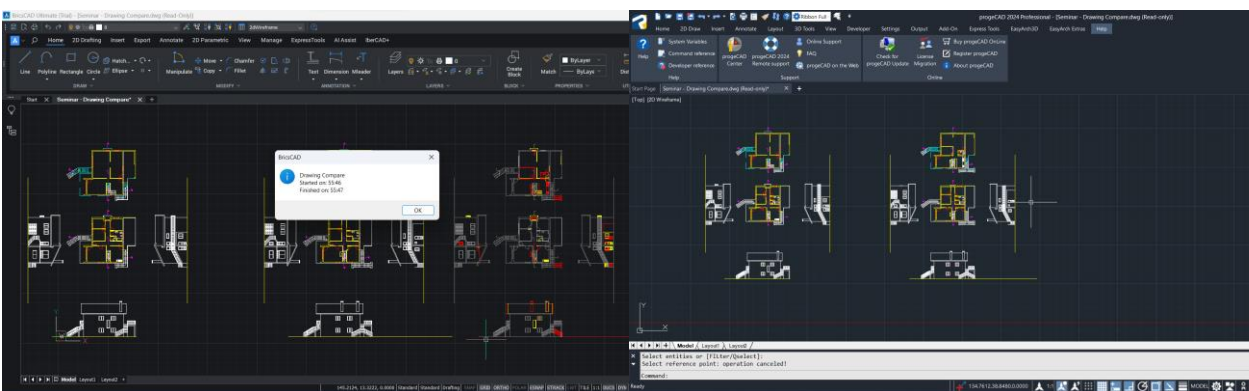


Figure 16: BricsCAD output after “Drawing Compare” AutoLISP script is run

Figure 17: ProgeCAD output after “Drawing Compare” AutoLISP script is run

4 Conclusion

It is evident that AutoCAD is more advanced and has an extensive set of tools for 2D as well as 3D drafting works. A detailed documentation is provided for drafters as well as developers looking to streamline jobs on AutoCAD. Furthermore, there are multiple forums

and communities of users online which makes help more accessible than any of the other three. However, for the drafting needs of EASE Engineering PLC, the price is too expensive. The unavailability of perpetual licenses makes matters worse.

Even though ProgeCAD is the cheapest option, it lacks documentation and when it comes to its APIs it did not perform the basic tasks that were selected as the basis of comparison. The most probable cause for the errors in the AutoLISP scripts is assumed to be differences in function names and syntactic differences. But the unavailability of documentation and in-built IDE has made debugging impossible.

ZWCAD and BricsCAD on the other hand are both lightweight, cheaper than AutoCAD, fast in execution and user-friendly. Even though ZWCAD under performed on the “Tributary Area” script, its price makes up for it. The author has also found the user experience to be the closest to AutoCAD than the rest.

In conclusion, it is recommended that EASE Engineering PLC, adopts ZWCAD for its drafting work which gives it most affordable and highest performance for the drafting needs.

References

- [1] S. Jin, Y. Zhang, T. Yamazaki, and Z. Jiang, “Automatic 3D CAD Model and 2D Drawings Generation in Construction Engineering,” *Journal of Physics: Conference Series*, vol. 1827, no. 1, p. 012115, Mar. 2021, doi: 10.1088/1742-6596/1827/1/012115.
- [2] L. Zhang and P. Zhang, “CAD secondary development technology based on. NET API,” *IOP Conference Series: Materials Science and Engineering*, vol. 768, no. 7, p. 072052, Mar. 2020, doi: 10.1088/1757-899x/768/7/072052.
- [3] A. Tzotzis, C. Garcia-Hernandez, J.-L. Huertas-Talon, D. Tzetzis, and P. Kyratsis, “Engineering applications using CAD based application programming interface,” *MATEC Web of Conferences*, vol. 94, p. 01011, 2017, doi: 10.1051/mateconf/20179401011.
- [4] I. A. Magomedov, A. M. Bagov, and A. A. Elmurzaev, “Comparative analysis of CAD software packages for engineering design,” *IOP Conference Series: Materials Science and Engineering*, vol. 1155, no. 1, p. 012083, Jun. 2021, doi: 10.1088/1757-899x/1155/1/012083.

Comparative Analysis of Citrix and VMware Horizon Virtual Desktop Infrastructure (VDI) Solutions in Organizational Environment

Dawit Tekle

HiLCoE , School of Computer Science and Technology
dawittek123@gmail.com

Fanthaun Bogale

HiLCoE , School of Computer Science and Technology
fantishb@gmail.com

Abstract:

The article discusses increasing reliance on Virtual Desktop Infrastructure (VDI) solutions within modern organizations necessitates a comprehensive understanding of the impact of various VDI platforms on organizational environments. Among the leading solutions, Citrix and VMware Horizon offer diverse features and functionalities. However, a significant gap remains in the research concerning a systematic comparison of these platforms and their effects on organizational performance. This research thesis aims to gather insights and experiences from IT managers, employees of the Ethiopian Roads Administration ICT departments, and other stakeholders regarding Citrix and VMware Horizon VDI solutions. The primary objective is to conduct a thorough qualitative and quantitative analysis of the impact of these VDI solutions on organizational environments. Key factors to be explored include the management consoles and tools preferred by IT administrators, the VDI solutions best equipped to meet the evolving demands of organizational environments, the effectiveness of provisioning common BYOD devices in both Citrix and VMware VDI environments, and the user-friendliness of each solution in terms of management and support.

The findings provide a comprehensive comparative analysis of Citrix and VMware Horizon, focusing on their implications for organizational environments. Insights gained from this study will enable organizations to make informed decisions when selecting a VDI solution that aligns with their specific needs, including deployment methods, performance, scalability, management capabilities, security, and user experience. Ultimately, this research will contribute to the existing knowledge base, addressing identified gaps in the literature and assisting organizations in optimizing their environments to enhance overall productivity and efficiency.

Keywords: VDI, Citrix, VMware Horizon, Organizational Environment

1. Introduction:

Virtual Desktop Infrastructure (VDI) is a technology visualizes desktop environments, allowing desktop operating system usually Microsoft Windows to operate and be managed from a data center [2]. The virtual desktop is transmitted over a network to a client device, enabling users to engage with the operating system and its applications as if they were installed locally. This client device can be a standard

PC, a thin client, or a mobile device. As Maukonen [8] pointed out that VDI enables users to access desktop environments hosted on centralized servers, thereby enhancing flexibility, security, and scalability. This approach not only streamlines IT operations but also supports remote work and collaboration, making it an essential component in modern enterprise architecture. As businesses increasingly adopt digital transformation strategies, understanding the intricacies of VDI becomes crucial for leveraging its full potential. Virtual Desktop Infrastructure (VDI) has emerged as a trans-formative technology in the realm of information technology, enabling organizations to deliver desktop environments to users on demand. As *Nair and Pramod* [11], delved that the architectural considerations necessary for effectively implementing VDI solutions, along with strategies for optimizing performance. Their comprehensive survey highlights the critical components of VDI systems, including resource management and scalability, which are essential for meeting the diverse needs of modern enterprises.

Digital transformation plays an important role in enabling companies to build sustainable supply chain capabilities and implement effective sustainability strategies. As the authors of Dong et al.[5], proved that a conceptual framework to explain how the adoption of these digital technologies can enable organizations to develop sustainable supply chain practices, increase their competitive advantage, and improve overall business performance. When comparing Citrix and VMware Horizon VDI solutions in organizational environment research, it is crucial to analyze key aspects such as performance, ease of deployment, scalability, user experience, and cost-effectiveness.

The article highlights the key benefits of adopting Virtual Desktop Infrastructure (VDI) in organizations: Furthermore, VDI simplifies IT management through centralized administration, making updates and configurations more efficient. This infrastructure also aids in business continuity, ensuring quick recovery from hardware failures or disasters due to centralized backups [23].

- **Improved Data Security and Compliance:** VDI centralizes data in secure data centers, reducing data breach risks and aiding compliance, especially in regulated industries.
- **Increased Workforce Flexibility and Mobility:** Employees can access their virtual desktops from any location and device, enhancing productivity and collaboration, particularly for remote work.
- **Centralized IT Management and Reduced Costs:** VDI streamlines IT management, allowing faster desktop provisioning and maintenance, which lowers the need for on-site hardware and support resources.
- **Enhanced User Experience and Productivity:** VDI offers a consistent, high-performance desktop environment, improving employee satisfaction and fostering efficient workflows.

2. Literature Review

The growing demand for flexible working environments has propelled the adoption of Virtual Desktop Infrastructure (VDI) solutions, with Citrix and VMware emerging as leading providers [32]. This literature review explores existing research on Citrix Virtual Apps and Desktops and VMware Horizon, highlighting key features, benefits, limitations, and comparative analyses.

Virtual Desktop Infrastructure (VDI) solutions offer significant flexibility and scalability for modern enterprises, enabling secure and efficient remote work capabilities. Deploying VDI solutions in the cloud provides enhanced scalability and cost efficiency but requires careful planning to optimize costs and ensure sustainability [6].

This paper [6] provided a detailed comparative analysis of leading cloud providers, such as Amazon Web Services (AWS), Microsoft Azure, and Citrix Cloud, focusing on cost optimization strategies and sustainable practices. The discussion integrates technical details, performance metrics, sustainability considerations, and future trends to provide a comprehensive guide for running cost-effective and environmentally friendly VDI solutions in the cloud.

2.1. Citrix VDI Solution Features

Citrix Environment is a comprehensive digital Environment solution that provides secure and unified access to a wide range of resources, including applications, desktops, and content, from any location and any device [42]. This solution aggregates and integrates Citrix Cloud services, enabling end-users, referred to as subscribers, to access all the resources available to them through a single, centralized platform.

Citrix Virtual Desktop Infrastructure (VDI) provides a secure, scalable, and high-performance platform for delivering virtual desktops to users across various environments. By utilizing centralized management and advanced virtualization technologies, Citrix VDI simplifies desktop deployment and administration while ensuring seamless access on any device or network. Key features such as the Flex-cast Management Architecture (FMA), HDX technology for enhanced user experience, and hybrid cloud support make Citrix VDI an optimal solution for enterprises seeking to improve productivity, security, and operational flexibility.

Delivery Controller: The central management component that tracks user sessions, brokers connections, authenticates users, and manages desktop application delivery. It communicates with hypervisors or other services to host the virtual machines.

Virtual Delivery Agent (VDA): Installed on physical or virtual machines to enable registration with the Controller and manage the connection between the machine and the user device. The VDA communicates with the Controller's Broker Service.

Citrix Environment App: Installed on user devices, it provides secure, self-service access to applications, desktops, and documents.

Citrix Director: A web-based tool for IT support teams to monitor the environment, troubleshoot issues, and provide user support. **Citrix License Server:** Manages the licensing for Citrix products used in the deployment.

2.2. VMware Horizon VDI Features

VMware Horizon is a comprehensive Virtual Desktop Infrastructure (VDI) solution designed to manage, deploy, and maintain virtual desktops and applications for end-users. It provides a robust infrastructure that supports remote access, flexibility, and productivity across various devices and locations.

VMware Horizon offers a rich set of features designed to enhance Virtual Desktop Infrastructure (VDI) environments by delivering secure, flexible, and scalable virtual desktops and applications. Key features include centralized management, allowing IT teams to efficiently deploy, monitor, and manage desktops and apps across private, hybrid, or multi-cloud environments using a unified cloud-based console.

VMware Horizon's infrastructure is designed to provide a scalable and secure platform for delivering virtual desktops and applications. The architecture includes key components such as the Horizon Connection Server, which acts as the central management and brokering component, authenticating users and facilitating connections to virtual resources. It supports advanced technologies like Instant Clone for rapid desktop provisioning, App Volumes for streamlined application delivery, and Dynamic Environment Manager for personalized user settings [35].

Kushwaha et al. [23] explore desktop virtualization, focusing on its transformative potential through Virtual Desktop Infrastructure (VDI). Key benefits include enhanced security via centralized data storage, reduced hardware costs from thin client use, and improved flexibility for remote work. However, challenges such as high initial setup costs, reliance on strong network infrastructure, and potential latency issues are significant. The authors call for further research to overcome these technical limitations and explore innovative VDI use cases across industries.

Ahmed et al. [3] highlight VDI's role in creating flexible learning environments, enabling students to access resources from various devices while promoting collaboration. Financial benefits include reduced hardware costs and simplified IT management, although challenges like robust infrastructure and data protection remain critical for successful implementation. The literature emphasizes VDI's potential to revolutionize education by enhancing accessibility and supporting personalized learning experiences. The authors in [43] outline advantages such as better resource utilization and improved management capabilities while noting drawbacks like performance challenges and setup complexity. Their cost analysis addresses initial investments and ongoing operational costs, examining how scalability affects performance and offering optimization techniques.

A study on VDI solutions for enterprises emphasizes flexibility and scalability in supporting remote work [6]. It discusses the benefits of deploying VDI in the cloud, including cost efficiency, while stressing the need for careful planning. Comparative analyses of providers like AWS and Microsoft Azure offer guidance on cost optimization and sustainable practices.

Gupta and Kumar et al. [14][7] explore various desktop virtualization approaches, highlighting benefits and challenges. Gupta's research identifies issues in resource allocation and performance. The literature advocates for comprehensive cost-benefit analyses to inform technology investments.

The impact of VDI on interaction and work performance is significant, allowing remote access from centralized servers while enhancing employee engagement and promoting sustainability through reduced commuting [33]. Lee et al. [30] emphasize user-centered design principles to improve user satisfaction and task efficiency in VDI environments. Finally, Fernandes et al. [27] examine how organizations collaborate

using VDI technologies, highlighting their role in driving innovation and economic growth.

3. Related Works

3.1. Comparative Analysis of VDI Technologies VMware, Citrix and AVD

Patil et al.'s [4] 2021 study, "A Comparative Study of Cloud Desktop Virtualization Technologies," offers an in-depth analysis of various cloud desktop virtualization solutions. The study compares key technologies, including VMware Horizon, Citrix Virtual Apps and Desktops, Microsoft Azure Virtual Desktop (AVD), and Amazon Work-spaces.

VMware Horizon is noted for its centralized management and robust scalability but is limited by high costs and complex setup.

Citrix excels in user experience and security but is also expensive and requires significant IT expertise. AVD is recognized for its cost-effectiveness and seamless integration with Microsoft services, although it has limited customization options. Amazon Work-spaces provides an easy setup and cost-effectiveness but offers less control over infrastructure.

The study evaluates these technologies based on scalability, security, cost, and user experience, concluding that organizations must carefully assess their specific needs to choose the most suitable solution, as each technology has unique strengths and weaknesses.

3.2. Comparative Study of Desktop Virtualization in Cloud Computing.

As Kwek et al.'s [9] proved in their related works analysis of various desktop virtualization solutions within cloud environments. This related work compares multiple technologies, evaluating their performance, scalability, security features, and cost-effectiveness. It highlighted how different virtualization approaches can cater to diverse organizational needs, such as remote access, resource management, and user experience. The authors also discuss the challenges associated with implementing these solutions, including network dependency and the complexity of management and provides insights into the strengths and weaknesses of various desktop virtualization technologies.

3.3. Employee Perceived Satisfaction in using Citrix Environment Technology

Al Fehaidi and Khan [32] conducted a study to evaluate employee satisfaction and performance with Citrix Workspace technology in a Saudi company, using the Technology Acceptance Model (TAM) as the theoretical framework.

Correlation analysis: The research focused on understanding employees' perceptions of the technology's ease of use, usefulness, and its impact on remote work productivity.

Findings revealed a strong positive correlation between user satisfaction and performance, with perceived ease of use and usefulness being critical factors influencing acceptance.

The study also highlighted that Citrix Workspace enhanced remote work

capabilities by providing secure access to applications and systems, thus improving overall employee productivity. However, it identified barriers such as limited familiarity with the technology and infrastructure challenges that could hinder its full adoption.

This research contributes to understanding how employee satisfaction with workplace technologies like Citrix Workspace can drive performance improvements in remote work settings.

3.4. Comparative Analysis of Desktop Virtualization Using VMware, Citrix

The author of [37] provided an in-depth evaluation of three popular desktop virtualization platforms VMware, Citrix, Hyper-V, and VirtualBox. The authors conduct a comparative analysis based on various criteria, including performance, scalability, ease of use, and feature sets.

Key findings highlight the strengths and weaknesses of each platform. VMware is noted for its robust performance and extensive feature set, making it suitable for enterprise environments. Hyper-V is praised for its integration with Windows environments and cost-effectiveness, while VirtualBox is recognized for its user-friendly interface and flexibility, particularly in smaller setups. The paper aims to guide organizations in selecting the most appropriate virtualization solution based on their specific needs and resources.

4. Methodology

4.1. General Research Methodology

This study employs a mixed-methods research design to comprehensively evaluate the capabilities and performance of Citrix and VMware Horizon VDI solutions. Following the framework established by Mishra and Alok [15], qualitative data will be collected through interviews and focus groups with IT administrators to gain insights into their preferences for management tools and the adaptability of each solution to organizational needs.

Quantitative data will be gathered through surveys to assess the effectiveness of BYOD provisioning and overall user satisfaction. The research will also reference the methodologies outlined by Takonaand Dubey and Kothari [17], which emphasize the importance of triangulating data sources for robust findings. As Gupta and Gupta [36], further support the use of mixed methods, highlighting their effectiveness in capturing diverse perspectives within complex research contexts.

4.2. Research Design

The study utilizes a cross-sectional survey design, allowing for the collection of data at a single point in time. This approach facilitates the examination of differences between user experiences with Citrix and VMware VDI solutions across various organizations.

The survey will be structured to include both closed-ended and open-ended questions, enabling the collection of quantitative data while also allowing for qualitative insights. This research employs a quantitative approach to evaluate the comparative effectiveness of Citrix and VMware Virtual Desktop Infrastructure (VDI) solutions in organizational environments.

To conduct a comprehensive comparative analysis of Citrix and VMware Horizon VDI solutions in the context of organizational environment, this research employed mixed- methods approach [41] in combining quantitative and qualitative research designs

The quantitative component involved the collection and statistical analysis [44] of performance metrics, scalability data, and user experience ratings to objectively evaluate the features and functionalities of the two VDI platforms.

The qualitative aspect will leverage in-depth technical interviews to gain a deeper understanding of deployment, management capabilities, security measures, and the overall impact of Citrix and VMware Horizon on the organizational environment.

- **Quantitative Component:** It involves collecting and statistically analyzing performance metrics, scalability data, and user experience ratings to objectively assess the features of both platforms.
- **Qualitative Component:** In-depth technical interviews will provide insights into deployment, management capabilities, security measures, and the overall organizational impact of the two VDI solutions. The qualitative aspect will leverage in-depth technical interviews to gain a deeper understanding of deployment, management capabilities, security measures, and the overall impact of Citrix and VMware Horizon on the organizational environment.

4.3. Research Sample Size Distribution

- ◆ For this study, the research employed a convenience sampling approach to recruit participants. Convenience sampling involves selecting participants who are readily available and willing to take part in the research [29].
- ◆ Convenience sampling involves selecting participants who are readily available and willing to take part in the research. This sampling method was deemed appropriate given the exploratory nature of the study and the challenges in accessing the target population.
- ◆ The research leveraged their personal and professional networks, as well as online platforms, to identify and invite 113 potential participants who met the inclusion criteria.

- ◆ Convenience allowed the researchers to efficiently gather a substantial amount of preliminary data to inform the development of future, more rigorous studies on this topic.

4.4. Research Data Collection Methods.

This research will use multiple data collection methods to analyze two VDI solutions:

- ✓ **Questionnaires:** Distributed via Google Forms and printed copies to ICT departments of the Ethiopian Roads Administration (ERA) and Ethiopian Civil Works Corporation (ECWC).
- ✓ **Interviews:** Conducting in-depth technical interviews to assess strengths and weaknesses.
- ✓ **Literature Review:** Reviewing case studies and success stories from Citrix and VMware Horizon.

Data analysis tools will include:

- **Spreadsheet Software:** Microsoft Excel and Google Sheets for basic organization and visualization.
- **Statistical Software:** Tools like Quirkos, MAXQDA, SPSS, and R for advanced statistical analyses.

5. Data Analysis and findings

5.1. Cronbach's Alpha Readability of the datasets.

Cronbach's Alpha, also known as coefficient alpha, is a statistical measure used to assess the internal consistency or reliability of a set of scale or test items. High Cronbach's Alpha value suggests a strong correlation (0.956) among the variables being measured, indicating a reliable relationship among them.

5.2. Technical Feature Set Related Evaluation

Using a questionnaire survey a methodological framework ensured that a rigorous and objective assessment, enabling a thorough understanding of how these solutions impact IT administration, user experience, and overall environment efficiency as implied in the organizational settings.

Citrix feature set Analysis - a feature set comparison utilizing a 5-point Likert scale ranging from very poor to excellent, each rating level represents a distinct evaluation of the features being assessed.

A rating of 1, indicating very poor, suggests that the feature set lacks essential functionalities, is inefficient, or significantly underperform compared to expectations. VMware feature set Analysis most the respondents evaluation tends to Very Good(3) this implies. "very good" would typically represent the highest level of satisfaction or agreement with the features being assessed but with less mean value compare to Citrix.

5.3. VDI solutions Feature Frequency of Usage Comparison

Based on respondents the frequency usage rating for Desktop Virtualization and

Provisioning in Citrix VDI is 3.91 out of 5. This suggests a relatively high level of usage and significance attributed to desktop virtualization and provisioning within the Citrix VDI setup. Frequency usage within VMware VDI solutions, highlights the significance of application virtualization, desktop virtualization, provisioning, remote access, and mobility features in the context of VMware VDI deployments. This analysis provided the significance of application virtualization, desktop virtualization, provisioning, remote access, and mobility features in the context of VMware VDI deployments.

5.4. VDI Challenges and Severity Comparison

For licensing cost, the severity ratings range from 1 to 5, with a mean of 3.91 and a standard deviation of 1.110. A moderate level of concern among users regarding the licensing cost associated with Citrix VDI solution. terms of latency or performance issues, the severity ratings range from 1 to 5, with a mean of 0.92 and a standard deviation of 0.925. Similarly, showed compatibility issues associated with VMware VDI solution received moderate severity ratings. The ratings range from 1 to 4, with a mean of 2.98 and a standard deviation of 0.788. This indicates that users perceive compatibility issues to be of moderate concern.

5.5. Overall Satisfaction on VDI Solutions

Dependent Variables; summarized the comparison between Citrix VDI and VMware VDI in terms of overall satisfaction on scalability and resource allocation, and overall satisfaction in accommodating organizational and user workloads. For Citrix VDI, the minimum satisfaction score is 3, the maximum is 5, with a mean of 4.08 and a standard deviation of 0.647 for both metrics. The data suggests that users have higher overall satisfaction with Citrix VDI compared to VMware VDI in terms of scalability, resource allocation, and accommodating organizational and user workloads

5.6. Correlation Analysis

Correlation between variables related to Citrix and VMware VDI. For instance, the research examine the relationship between user experience and performance metrics or the correlation between cost factors and satisfaction levels results of the correlation analysis will be discussed in detail in next results and discussion chapter of research.

5.7. Theme Categorization and Coding

Coding is fundamental step in qualitative data analysis that involves systematically assigning labels or codes to segments of data. The research identifies patterns, themes, or concepts within the data and assigns relevant codes to capture these elements.

6. Results and Discussion

6.1. Demographic Data Summary

The demographic respondent statistics the sample consists primarily of male respondents (80%), with an average age between 36-45 years old (61%), a mix of bachelor's and master's degree holders (55% and 45% respectively), mostly mid-level to senior-level professionals (76%), from the same organization (67%), and with

an average work experience of 6-10 years (40%).

6.2. VDI Features Comparison Findings Summary

It offers a high degree of scalability, enabling organizations to easily expand their VDI deployments. Also in figure 6.1 result show that the collaboration and sharing capabilities being the most positively perceived (78.4%, 74.4%, and 72% respectively).

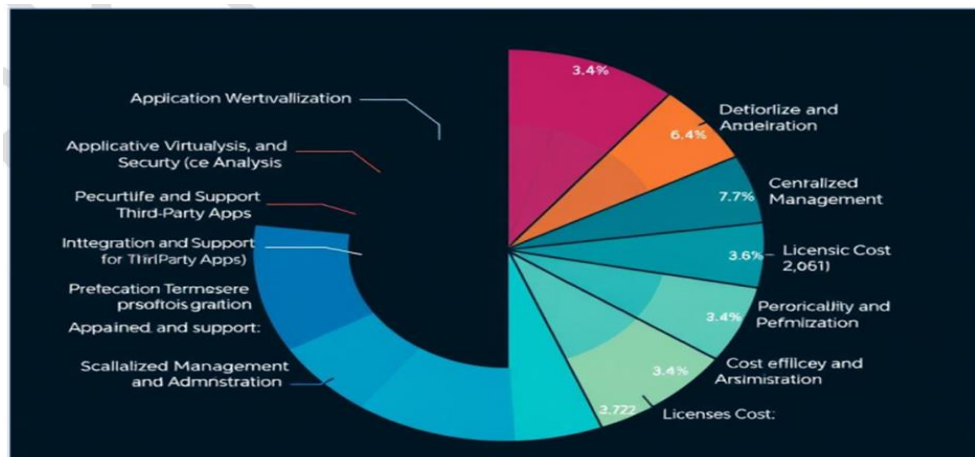


Figure 6.1 VDI Feature set Comparison Result

The licensing cost and the cost and efficiency and resource optimization features are perceived to be the least effective (52.2%). VMware VDI solution is perceived to be moderately effective across the cost efficiency and resource optimization, and the identified cost savings being the most positively perceived (77%).

6.3. VDI Challenges Comparison

In figure 6.2 above the Citrix solution is generally perceived to have higher licensing costs (78.2%), lower latency or performance issues (18.4%), and fewer compatibility issues (4.6%), while the VMware Horizon solution is perceived to have moderate flexibility and remote access capabilities (61.6%), some latency or performance issues (41.8%), and moderate compatibility issues (59.6%).

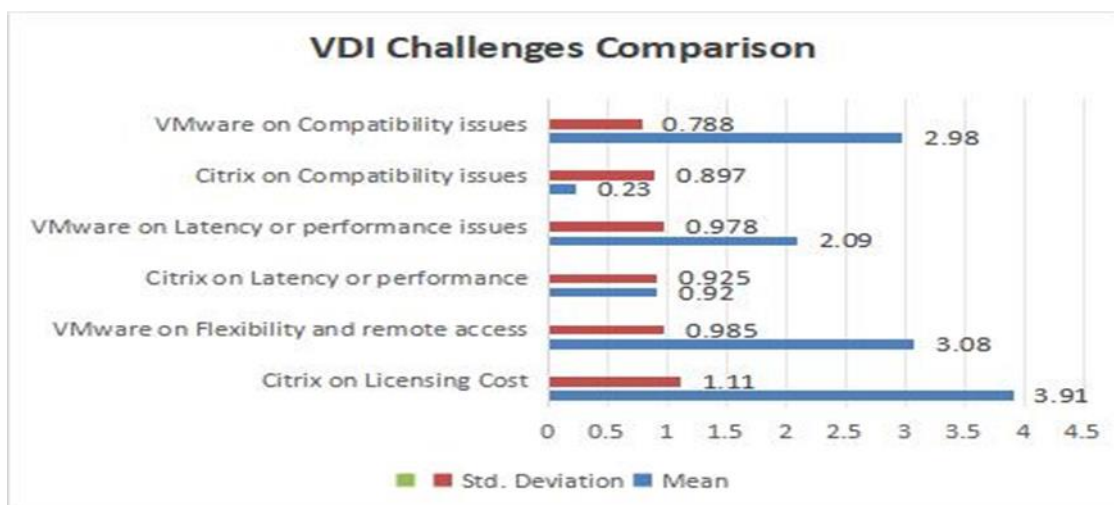


Figure 6.2 VDI challenges comparison

6.4. Correlation Analysis Result Summary

The correlation analysis also indicated that Level of education and Citrix VDI solutions users: Pearson Correlation: 0.856 (between level of education and Citrix VDI

solutions users) Sig. (2-tailed): 0.986 The correlation coefficient between level of education and Citrix VDI solutions users the level of education and Citrix VDI solutions users is 0.856, indicating a strong positive correlation.

The correlation coefficient of 0.856 suggests a relatively strong linear relationship between level of education and Citrix VDI solutions users the two variables.

The Kendall's tau_b correlation coefficient between level of education and Citrix VDI solutions users the level of education and Citrix VDI solutions users is 0.856, indicating a relatively strong positive correlation.

6.5. Overall Satisfaction Comparison on VDI Solutions

In **Figure 6.3** the data suggests that organizations are more satisfied with the Citrix VDI solution's scalability, resource allocation, and ability to accommodate their workloads, with an 81.6% satisfaction level, compared to the VMware VDI solution, which has a 61.8% satisfaction level.

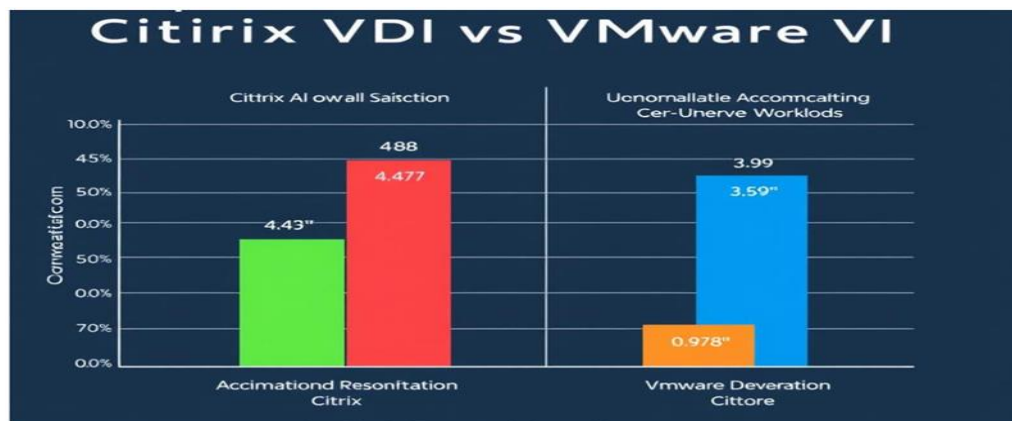


Figure 6.3 overall satisfaction comparison on VDI

6.6. Results of Thematic Analysis

This thematic analysis, which conducted a comparative evaluation of Citrix and VMware Horizon VDI solutions, reveals several key themes that organizations must weigh when selecting the optimal virtual desktop infrastructure for their needs. The analysis was based on 15 semi-structured interviews conducted between December 2023 and February 2024, which identified 10 major themes.

❖ Responsibilities within the Organization

ERA: Participants included IT managers and technical staff, emphasizing roles focused on technology evaluation and implementation. This suggests a strong technical expertise within the organization.

ECWC: Participants consisted of a mix of IT and operational leaders, indicating a broader perspective on technology's impact on organizational performance and employee experience.

❖ Prior Experience with VDI/Remote Work

ERA: Many participants reported extensive prior experience with VDI solutions, indicating familiarity with various deployment models and a proactive approach to

technology adoption.

ECWC: Participants exhibited a range of experiences, with some having limited exposure to VDI prior to implementation. This variability suggests a learning curve for some employees as the organization embraced remote work.

❖ *Perceptions of VDI Solutions*

ERA: Participants perceived Citrix VDI as a superior solution, praising its integration capabilities and robust features. The focus was on performance and adaptability to organizational needs.

ECWC: While expressing satisfaction with VMware, ECWC participants highlighted user experience and performance as critical factors. The perception here leans towards evaluating solutions based on end-user impact and operational efficiency.

❖ *VDI Implementation Process*

ERA: The implementation process was characterized by strong leadership involvement, with IT managers taking a central role in evaluating and selecting VDI solutions. This collaborative approach facilitated a smoother transition.

ECWC: Participants noted a more decentralized approach, with varied roles contributing to the implementation process. This suggests a need for cross-departmental collaboration but may also reflect differing levels of involvement and expertise among staff.

❖ *Organizational Factors*

ERA: The organization's size and technology strategy were cited as key factors influencing VDI adoption. A clear alignment between VDI solutions and organizational goals was evident, supporting effective decision-making.

ECWC: Factors such as industry-specific challenges and existing IT infrastructure played a significant role in shaping VDI adoption. The organization's focus on community engagement and social responsibility also influenced the selection of technology solutions.

The thematic analysis highlighted distinct differences between ERA and ECWC in their experiences and perceptions of VDI solutions. ERA's approach is characterized by strong technical expertise and a focus on integration and performance, while ECWC emphasizes user experience and collaborative implementation. Understanding these organizational nuances is essential for tailoring VDI solutions that meet the unique needs and goals of each organization, ultimately enhancing satisfaction and effectiveness in remote and hybrid work environments.

❖ *Benefits of VDI in Organizational Environment*

The thematic analysis revealed key goals and objectives among participants regarding the implementation of VDI solutions. These objectives were primarily driven by the need to address specific business challenges and improve organizational performance. **Enhanced Remote Work Capabilities:** A significant goal identified was the desire to improve remote and flexible work capabilities for employees. Participants emphasized the importance of providing a consistent and reliable

experience for remote workers to maintain productivity and engagement.

Improved Accessibility and Mobility: Participants noted that enhancing end-user accessibility and mobility was critical. The VDI solution was seen as a means to enable employees to access their work environments seamlessly, regardless of location, thereby supporting a more flexible work culture.

User Experience Focus: There was a strong emphasis on improving user experience. Participants expressed the need for a VDI solution that could deliver high performance and responsiveness, which is essential for maintaining employee satisfaction and productivity.

Increased Productivity and Operational Efficiency: Many participants cited the goal of increasing employee productivity and operational efficiency. By modernizing their IT infrastructure through VDI implementation, organizations aimed to streamline operations and reduce the time and resources needed for IT management. The analysis of goals and objectives for VDI implementation reveals a strong focus on enhancing remote work capabilities, improving user experience, and increasing operational efficiency. Participants from both ERA and ECWC articulated their needs for a modern, scalable VDI solution that aligns with their strategic objectives.

❖ *Based on Performance and Scalability VDI Environment.*

Based on the thematic analysis and participant feedback regarding performance and scalability in VDI environments.

ERA Organization

Performance: Participants from ERA reported high satisfaction levels with Citrix VDI, highlighting its robust performance in delivering a consistent user experience across remote and hybrid work settings. The solution was praised for its ability to handle high workloads and provide seamless access to applications.

Scalability: Citrix VDI was viewed as highly scalable, allowing ERA to easily adapt to changing workforce demands and rapidly provision virtual desktops for new employees or temporary staff. The integration with existing infrastructure also facilitated smooth scaling.

ECWC Organization

Performance: While participants acknowledged the strengths of VMware VDI, there were mixed reviews regarding its performance compared to Citrix. Some noted that VMware offered good overall performance, but it did not match the high satisfaction levels reported for Citrix VDI.

Scalability: VMware VDI was mentioned as capable of supporting the organization's growth. However, some participants indicated challenges in scaling the legacy infrastructure effectively, which impacted their experience with VMware.

In summary, the ERA organization, utilizing Citrix VDI, excels in both performance and scalability. Participants expressed strong satisfaction with the solution's ability to support remote work and adapt to organizational needs.

In contrast, while ECWC uses VMware VDI effectively, it does not achieve the same level of satisfaction regarding performance and scalability as noted by participants from ERA. Organizations aiming for high performance and scalability in their VDI environments may prefer Citrix VDI based on this analysis.

Security and Authentication Methods

Based on the thematic analysis and participant feedback regarding security and authentication methods in VDI environments.

ERA Organization on Citrix VDI Solution:

Security: Participants from ERA highlighted the strong security features of Citrix VDI, including advanced encryption, secure access controls, and comprehensive monitoring. The solution's ability to protect sensitive data while providing remote access was a significant advantage.

Authentication Methods: Citrix VDI supports multi-factor authentication (MFA) and integration with various identity management solutions, enhancing security for remote users. Participants appreciated the flexibility in authentication options, which aligns with organizational security policies.

ECWC Organization on VMware VDI Solutions

Security: While VMware VDI also offers solid security features, some participants noted that their experience with legacy systems posed challenges in implementing the latest security protocols. Overall, VMware VDI provides essential security measures, but there were concerns regarding some aspects of data protection and compliance.

Authentication Methods: VMware VDI supports various authentication methods, including SSO (Single Sign-On) and MFA. However, some participants indicated that the integration of these methods with existing systems could be complex, impacting their effectiveness.

6.7. Discussion

The findings of this research align with the comparative insights provided by Patil et al. [4]. Citrix VDI solutions are recognized for superior performance, characterized by lower latency and fewer compatibility issues, thanks to advanced features like HDX protocols and robust integration with hypervisors and cloud platforms (e.g., AWS, Azure, Google Cloud). This supports Patil et al.'s observation that Citrix excels in delivering a seamless user experience and efficient resource utilization, making it ideal for high-performance and flexible environments. In contrast, VMware Horizon offers moderate flexibility and remote access but experiences occasional latency and compatibility issues due to its reliance on proprietary ecosystems like VMware ESXi. Patil et al. [4] highlight that while VMware integrates well within its ecosystem, this vendor lock-in can limit adaptability in multi-cloud or hybrid environments. Citrix's support for multiple hypervisors and advanced provisioning services enhances its performance scalability for large deployments. However, Citrix typically comes with higher licensing costs compared to VMware Horizon, which offers competitive pricing models. Cost considerations are critical in selecting VDI

solutions, with Citrix's higher costs potentially deterring smaller organizations. Yet, its superior performance makes it a preferred choice for enterprises prioritizing reliability and user experience. Conversely, VMware Horizon's cost-effectiveness and integration with existing VMware technologies make it practical for those already invested in that ecosystem. In conclusion, Citrix is better suited for organizations seeking high performance and flexibility across multi-cloud environments, albeit at a higher cost. VMware Horizon provides a balanced solution for those looking for moderate flexibility and cost efficiency within a proprietary framework. This analysis reinforces Patil et al.'s emphasis on aligning VDI selection with organizational priorities such as cost constraints, scalability needs, and performance requirements.

7. Conclusion and Recommendation

The use of virtual desktop infrastructure (VDI) solutions is increasingly common among organizations seeking to enhance efficiency, security, flexibility, and cost savings. By centralizing management, improving security, facilitating mobility, and optimizing resource utilization, VDI has become a valuable tool for businesses with modern, distributed workforces. This comparative analysis aimed to identify key factors for organizations when selecting a VDI solution to meet their specific needs. The thematic analysis highlighted drivers of VDI adoption, such as improved remote work capabilities, enhanced end-user productivity, and enabled IT modernization. Organizational context—such as size, industry, and culture—significantly influences VDI implementation and outcomes. Participants noted both benefits and challenges, including cost, performance, compatibility, and security considerations.

Based on the findings, several actionable recommendations for ERA and ECWC are proposed:

Assess Organizational Needs: Conduct a thorough evaluation of IT infrastructure, user requirements, and operational goals to identify the most suitable VDI solution. For ERA, this may include assessing the need for high-performance computing for data analysis, while ECWC might prioritize secure access to sensitive conservation data.

Evaluate Cost-Benefit Analysis: Consider both short-term costs (e.g., licensing fees) and long-term benefits (e.g., reduced hardware expenses, improved productivity) when selecting a VDI platform. ERA and ECWC should balance these factors against their budget constraints and operational priorities.

Invest in Training and Support: Provide adequate training for IT staff and end-users to maximize the benefits of VDI solutions and address compatibility or performance issues. **Monitor and Adapt:** Continuously assess user satisfaction and system performance, adjusting. Ongoing monitoring and optimization of VDI performance, security, and user experience are essential to maximizing benefits.

In the future, further research is needed to explore the long-term impacts of VDI on employee satisfaction, productivity, and organizational performance.

Cloud-Native VDI Solutions: Explore the potential of cloud-native VDI

solutions to offer greater scalability and cost efficiency compared to traditional on-premises deployments, which could be beneficial for ERA's remote fieldwork and ECWC's collaborative projects.

User Experience and Satisfaction: Conduct longitudinal studies to assess how VDI solutions impact user satisfaction and productivity over time, identifying areas for improvement. This would be particularly valuable for ERA and ECWC, where user experience directly affects research outcomes and conservation efforts.

References

- [1] KIM, Tae-Hi. "How to Implement Successful Virtual Desktop Infrastructure (VDI) in the Manufacturing Sector." *The Journal of Industrial Distribution & Business* 13.10 (2022): 15-22.
- [2] Mixon, E., & Harder, J. (n.d.). What is virtual desktop infrastructure? VDI explained. TechTarget.<https://www.techtarget.com/searchvirtualdesktop/definition/virtual-desktop-infrastructure-VDI>
- [3] Ahmed, Mariwan, et al. "Enhancing Educational Paradigms: A Comprehensive Review of Virtual Desktop Infrastructure (VDI) Applications in Learning Environments." *The Indonesian Journal of Computer Science* 13.2 (2024).
- [4] Patil, A. S., et al. (2021). "A Comparative Study of Cloud Desktop Virtualization Technologies."
- [5] Dong, L., Hao, S., Ren, J., & Zhang, L. (2022). Toward an understanding of the impact of cloud computing on organizational performance. *Information & Management*, 59(1), 103-460.
- [6] Thumala, Srinivasa Rao. (2021). *Running Sustainable Virtual Desktop Infrastructure (VDI) Solutions in the Cloud*.
- [7] Gupta, Ambika. (2023). An approach for desktop virtualization and challenges using cloud sim." *AIP Conference Proceedings*. Vol. 2721. No. 1. AIP Publishing, .
- [8] Maukonen, Henna-Riikka. "Virtual Desktop Infrastructure." (2021).
- [9] Kwek, K. Y., et al. (2020). A Comparative Study of Desktop Virtualization in Cloud Computing.
- [10] Stutz, Andreas, et al. "Automated Integration of Remote Terminal Units via IEC Protocol with the Module Type Package." 2022 IEEE 27th International Conference on Emerging Technologies and Factory Automation (ETFA). IEEE, 2022.
- [11] Nair, J., & Pramod, K. V. (2021). Virtual desktop infrastructure: Architectural considerations and performance optimization. *ACM Computing Surveys*, 54(6), 1-35. 16. Patel, A., & Reddy, A.
- [12] Attaran, Mohsen, Sharmin Attaran, Attaran, Mohsen, Sharmin Attaran, and Diane Kirkland. "The need for digital workplace: Increasing workforce productivity information age." *International Journal of Enterprise Information Systems (IJEIS)* 15.1 (2019): 1-23.(12)
- [13] Sekwakwa, Matthews, and Sello Mokwena. "Case Notes: Factors Influencing the Adoption of Virtual Desktop Infrastructure (VDI) Within the South African Banking Sector." *The African Journal of Information and Communication (AJIC)* 15 (2019)
- [14] Rupesh Kumar, Arun Kumar Yadav, H N Verma, "An analysis of Approaches for Desktop Virtualization and Challenges", *International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT)*, ISSN:2456-3307, Volume 7 Issue 4, pp. 600-612, July-August 2021.
- [15] Mishra, S. B., & Alok, S. (2022). *Handbook of research methodology*. Educational publishing.
- [16] McGarrigle, Malachy. McGarrigle, Malachy. "Watch points for consideration when utilizing a VDI network to teach Archicad BIM software within an educational programme." (2023).(16)

- [17] Takona, J. P. (2024). Research design: qualitative, quantitative, and mixed methods approaches. *Quality & Quantity*, 58(1), 1011-1013.
- [18] Chen, Y., Li, S., Li, C., Li, Y., Li, Z., & Wang, L. (2021). An Evaluation of Virtual Desktop Infrastructure Performance and Scalability. *IEEE*
- [19] Hu, Yunfei. "Application of VDI Virtual Desktop Cloud-Based Computer Room Construction in Universities." *Journal of Computer Technology and Electronic Research*, vol. 1, no. 1, Aug. 2024, <https://doi.org/10.70767/jcter.v1i1.66>.(19)
- [20] Auinger, Andreas, and Werner Wetzlinger. "Prohibiting Bring Your Own Device (BYOD) in Companies: Effectiveness and Efficiency vs. Satisfaction." *HCI in Business, Government and Organizations. Information Systems and Analytics: 6th International Conference, HCIBGO 2019, Held as Part of the 21st HCI International Conference, HCII 2019, Orlando, FL, USA, July 26-31, 2019, Proceedings, Part II 21*. Springer International Publishing, 2019.11.
- [21] Wang, S., Liu, S., & Fang, J. (2020). A Cost-Benefit Analysis of Desktop Virtualization in Higher Education. *Journal of Information Technology Education: Research*, 19, 197-216.
- [22] Andrews, Scott. *Learning and Researching with Case Studies: A Student Companion for Business and Management Research*. Routledge, 2023.
- [23] Kushwaha, S., Yadav, A. K., & Verma, H. N. (2023). Desktop Virtualization: Benefits, Challenges, and Future Trends. *International Journal of Education and Management Engineering*, 13(6), 14.
- [24] Sagi, Sriramraju. "Revolutionizing Remote Work Through Virtual Desktop Infrastructure: A Comprehensive Analysis of VMware Horizon." *Journal of Artificial Intelligence*; June 2023, pp 1–4, [https://doi.org/10.47363/jaicc/2023\(2\)227](https://doi.org/10.47363/jaicc/2023(2)227).
- [25] A. Rahman Alzoubaidi, M. Alzoubaidi, I. Abu Mahfouz, T. Alkhamis, F. Fuad Salim AL-Asali, and M. Alzoubaidi, "Virtual desktop infrastructure in higher education institution: an application of home and mobile computing environment," *Azerbaijan Journal of High-Performance Computing*, vol. 4, no. 1, pp. 29–38, Jun. 2021, doi: 10.32010/26166127.2021.4.1.29.38.
- [26] Sagi, Sriram. "Enhancing Virtual Desktop Infrastructure with Hybrid Cloud Environments and GPU Acceleration." *Journal of Artificial Intelligence & Cloud Computing*, June 2023, pp. 1–5, [https://doi.org/10.47363/jaicc/2023\(2\)211](https://doi.org/10.47363/jaicc/2023(2)211).
- [27] Fernandes, C., Ferreira, J. J., & Marques, C. S. (2019). Technology transfer between university and public and private enterprises. *Journal of Engineering and Technology Management*, 51, 45-65.
- [28] Shad, Ben. "Implementing a virtual desktop infrastructure at the District of North Vancouver." (2020).
- [29] Etikan, Ilker, Sulaiman Abubakar Musa, and Rukayya Sunusi Alkassim. "Comparison of convenience sampling and purposive sampling." *American journal of theoretical and applied statistics* 5.1 (2016): 1-4.
- [30] Lee, S., Kim, S., Choi, S., & Shin, Y. (2020). A User-Centered Design Framework for Desktop Virtualization. *Journal of Usability Studies*, 15(4), 164-179. Volume 13 (2023), Issue 6
- [31] VMware, VDI: A New Desktop Strategy, https://www.vmware.com/pdf/vdi_strategy.pdf.
- [32] Al Fehaidi, Bandar Ali, and Muhammad Asif Khan. "Towards Employee Perceived Satisfaction in using Citrix Environment Technology." *International Journal of Advanced Computer Science and Applications* 13, no. 4 (2022).
- [33] Kwak, Young, and Min Soo Shin. "The Effect of VDI Technical Characteristics on Interaction and Work Performance." *Journal of Information Technology Services* 20.4 (2021): 95-111.
- [34] Dubey, U. K. B., & Kothari, D. P. (2022). *Research methodology: Techniques and trends*. Chapman and Hall/CRC.
- [35] VMware Horizon outperforms Citrix in Principled Technologies report -VMware End-User Computing Blog

- [36] Gupta, A., & Gupta, N. (2022). Research methodology. SBPD publications.
- [37] Anwar, M. A., et al. "Comparative Analysis of Desktop Virtualization Using VMware, Hyper-V and VirtualBox." (2020).
- [38] Liu, L., Zhang, J., & Wang, Q. (2021). An Energy-Efficient Desktop Virtualization Architecture. *Journal of Computer and System Sciences*, 118, 30-46.
- [39] Zoakei, R. B., et al. "A Comparative Study of Desktop Virtualization Solutions." (2018).
- [40] Knobbout, Jan Jaap, and Z. A. Lubsen. "Virtualized Simulation for Military Concept Development and Experimentation: The Cerebro Battle Lab, a Case Study." I/ITSEC, 2023.
- [41] Yin, Robert K. Case study research and applications. Vol. 6. Thousand Oaks, CA: Sage, 2018.
- [42] <https://www.citrix.com/blogs/2023/05/03/level-up-your-citeworkspaceenvironments-with-a-new-ui/>
- [43] Rot, Artur, and Pawel Chrobak. "Benefits, Limitations and Costs of IT Infrastructure Virtualization in the Academic Environment. Case Study using VDI Technology." ICSOFT. 2018.
- [44] Golzar, Jawad, Shagofah N. and Omid Tajik. "Convenience sampling." *International Journal of Education & Language Studies* 1.2 (2022): 72-77.

Practicing ESB Technology for Internet, Mobile Banking and Third Party Services Integration in case of Tsehay Bank

Baheredin Surur
HiLCoE School of Computer Science & Technology
babasurr@gmail.com

Abstract

This paper discusses the importance of enterprise service bus (ESB) in facilitating service integration and meeting the changing demands of businesses, particularly in financial institutions. With many banks in Ethiopia adopting ESB, it is crucial for new players like Tsehay Bank to choose the right, efficient, and cost-effective ESB. The paper presents a comparative analysis of four widely used ESB technologies - WS02, Red Hat Fuse, Fiorano, and Apache ServiceMix and recommends the best ESB option for Tsehay Bank. Overall, this paper highlights the benefits of employing ESB in service integration and recommend one of the ESB tool mentioned above for Tsehay bank digital banking division.

Keywords: Integration, API, Gateway, tools, Enterprise service bus (ESB), applications, classification, service oriented architecture (SOA)

1. Introduction

Tsehay Bank is one of the banks that has recently joined the banking industry in Ethiopia. [12] Tsehay Bank is a share company established and started giving services for its customers on July 23, 2022 with a business license issued as per the banking business proclamation No 592/2008. It commenced operations with an authorized capital of Birr 2.8 billion and paid-up capital of Birr 734 million. The Bank's paid-up capital has reached Birr 1.1 billion (as of August 7, 2023) and raised its subscribed capital to Birr 5 billion.

The bank currently has over 750 employees. Tsehay Bank celebrated its one-year anniversary a few months ago and during this short period of time, the bank was able to open more than 80 branches for conventional banking service and started offering digital banking services such as mobile banking, ATMs, telegram bot services and more by utilizing Temenos core banking solution and Oracle database. Overall, Tsehay Bank has made significant progress in its first year of operation, but there is still room for improvement. By investing in infrastructure, talent, and customer experience, the bank can establish itself as a strong player in the Ethiopian banking industry by aspiring to deliver sustainable values for stakeholders with best customer experience and through state-of-the-art technology.

Tsehay Bank, despite being new to the banking industry, has already started offering

digital banking services such as payment cards, mobile and online platforms. The bank aims to achieve accessibility, process efficiency, and improved competitiveness by managing its existing system development process and infrastructure using an agile iterative approach. Additionally, the bank is waiting for the main digital and infrastructure project to commence to make its services robust enough to meet the ever increasing demand of customers with changing technology and business model and regulation.

Tsehay Bank also completed a lot of third party integration with Ethiotelcom, Ethswitch, Safarcom, MPESA, DARs, ECX, NBE, Guzo Go and others many more in the pipe line. Being new in the industry has not limited the bank to provide most of the services provided by other pioneers in the business. To expand the currently active services and incorporate the emergent services the bank still need to do a lot in infrastructure development and acquiring right tool for integration of different in house and third party applications. Used in House developed middleware (using Java spring boot framework) which lacks micro-service capability which make change and scalability difficult and any change made in one services will affect all system integrated in the middleware used. With current middleware used third party integration is not with required time frame and difficulty in deployment, protocol conversion, Micro services, plug and play features and scalability, flexibility with interface and automated test

Due to globalization and digitization most business operations are continuously changing and integration with different companies are becoming a necessity, and when it comes to a bank the stake is beyond that. Because of its emergent nature it is crucial to have well regard open application integration tools that can facilitate third party integration with different connection and format modalities. The concept of an Enterprise Service Bus (ESB) emerged in the early 2000s as a response to the challenges posed by traditional point-to-point integration methods. During that time, businesses faced significant costs and maintenance issues with their point-to-point connections for system integration.

This practice can be very helpful for Tsehay Bank to bring new reliable services quickly for customers. Enterprises all over the world have to face a fierce competition. In order to stay in business, they constantly have to automate their business processes, integrate with their business partners and provide new services to their customers.

Thus, the objective of this paper is to recommend the best Enterprise Service Bus (ESB) tools to Tsehay Bank and support in choosing the one that is good enough in facilitating communication and integration of applications and services.

2. Related Work

2.1 Application Integration Progress and ESB Technologies

Harmonizing to [3] IT infrastructure can have a wide range of diverse technologies, platforms and applications that need to interoperate to serve business goals. Furthermore [3]

explained Enterprise Application integration (EAI) which can be defined as “the enablement of data sharing between heterogeneous information systems within one enterprise” is becoming more and more essential to enterprises as it streamlines business processes, ensures the integrity between different systems, reduces labor cost and minimizes human errors. It is mentioned on [3] IT infrastructure runs a variety of diverse technologies, platforms and systems that may involve legacy systems, ERP solutions, and cloud-based systems, makes such goal hard to achieve, especially with the rapid growth of information technologies and non-stop business growth that rely heavily on Information Technology. Organizations like banking industry (Fintech) that relay on IT for their day today services are required to merge the different IT components and solutions together to form a solid infrastructure that enables easy sharing of data and information across different systems in a seamless fashion.

According to [2] the first generation of EAI involves the use of point-to-point middleware which provides APIs or program interfaces to link one application to another using message queuing technology such as Common Object Request Broker Architecture (CORBA) or remote procedure calls (RPCs). Here the middleware feature is embedded into applications and need extra work or modification at both end to complete the integration process and also it is kind of one to one relationship and it is difficult to manage as the number of integration increase and correspondingly increase development cost, maintenance and change.

The second generation of EAI is the centralized integration platforms which is known as message broker. In this integration approach, the message broker is a central hub that is responsible for data transformation and routing. Here the process of connecting various systems and transferring messages between them can be a complex task. However, with the use of adapters, this process is made simpler and more efficient. Adapters are used to connect all systems to a hub, allowing for seamless communication between them. In addition to facilitating communication, adapters also play a crucial role in converting data formats into a format that the hub can understand. Once the sender application publishes messages to a message broker, the message broker takes over and processes the message by applying data transformation. This ensures that the message is delivered in the correct format and can be understood by all systems involved in the communication process. Overall, the use of adapters and message brokers is essential for ensuring smooth and efficient communication between different systems.

In recent years, EAI moved toward loosely coupled fashion and standard exchange semantics. This can be seen in Enterprise Service Bus (ESB) integration technology which is standard-based integration platform that emerged Service Oriented Architecture (SOA) technologies and Message Oriented Middleware (MOM) to standardize information flow and message processing using a central hub [4]. ESB is a highly distributed messaging backbone which combines traditional middleware technology and web services to towards

more loose coupling and standard semantics. ESB provides flexibility, reusability and neutrality of SOA technology and web services to the integration world. The fact that SOA technology is platform- and programming language- independent, makes it ideal solution to unify the integration access diverse systems [5].

As stated on [11] architecturally in the absence of ESB products, heterogeneous applications could be integrated through alternative architecture as shown in figure 1. However, this application integration architecture is difficult to scale and maintain. Moreover, adding an application would mean interface customization of all other networking applications.

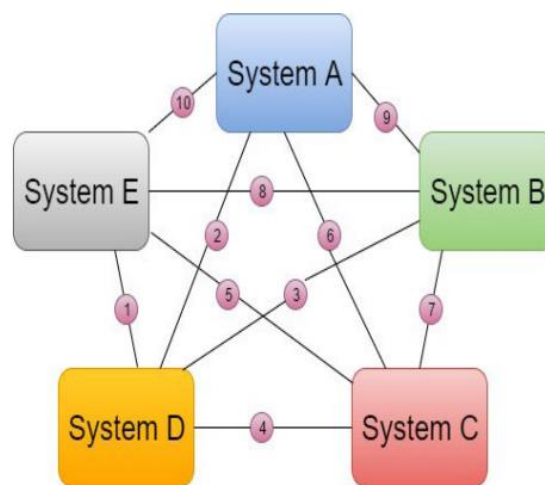


FIGURE 1. Point to point integration architecture.

Enterprise service bus architecture: As shown in figure 2, applications are indirectly linked through ESB instead of being linked directly to each other. Enterprise service bus architecture. Responsible for all the built-in logic necessary to make the structures interact/integrate.

The ESB was introduced as a more efficient and cost-effective solution for integrating systems and applications. Because an ESB controls the way work moves, it makes it easy to change components or add additional components to an application. It also makes for a convenient place to enforce security and compliance requirements, log normal or exception conditions and even handle transaction performance monitoring. Finally, an ESB also provides load balancing, in which multiple copies of a component can be instantiated to improve performance, as well as failover support should a component or resource fail.

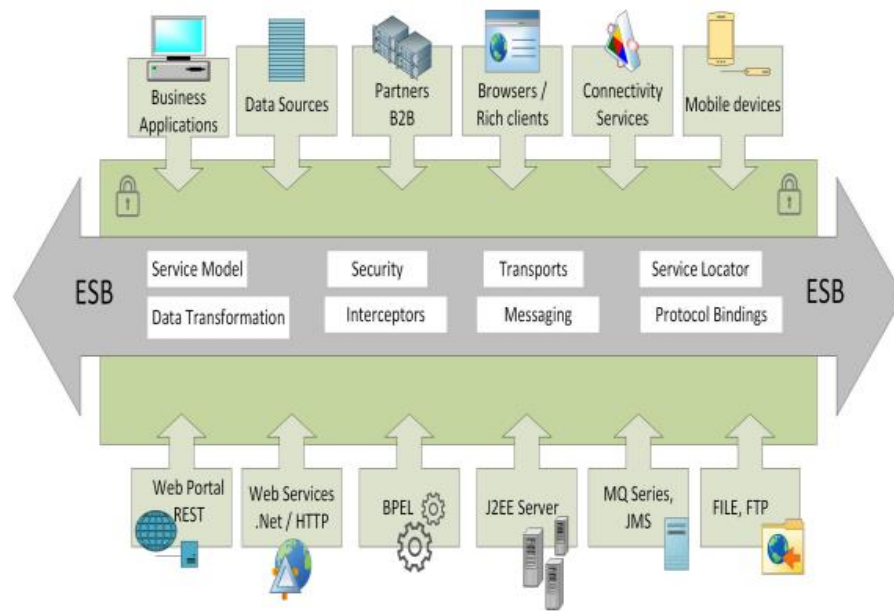


FIGURE 2. Enterprise service bus architecture.

2.2 Benefits of ESB

There are many advantages to us to get by including ESB in the integration of different applications and services

ESB enables Standardized integration: It reduces proprietary vendor lock-in and increases the ability to integrate with business partners using standard interface and protocols. It also helps businesses in leveraging the existing IT staff instead of looking talent or specialized integration architects outside.

ESB is loosely-coupled: This is also essential aspect of ESB as the application logic is separated from the connection infrastructure and independent web services can be accessed as shared services by the distributed systems connected to the bus which grants more flexibility and scalability

ESB will improve services flexibility and reusability: A pattern which uses web services, intelligent routing and message transformation and services orchestration in order to make services more flexible and reusable. Thus, the company is not just prepared for the short run but also for the long run, for the future technological changes and evolution.

ESB enables Service substitution feature: This enable the substitution of one service implementation by another with no effect to the clients of that service. This requires both the service interfaces that are specified in a manner that is independent of the service location and communication protocol that is involved.

ESBs can be easily scaled horizontally by adding more instances to meet growing

demands. This enables organizations to handle large volumes of data and transactions efficiently.

ESB offers flexibility in terms of message routing and transformation. It supports many protocols and message formats and can transform messages between different formats. Organizations can then integrate diverse messaging applications with ease.

ESB supports various security mechanisms, such as authentication, authorization, and encryption, which ensure the confidentiality and integrity of the data.

ESBs provide a library of reusable components that can be leveraged across different applications and systems. This reduces development time and effort, giving organizations greater code reusability.

ESB enables support of versatility: is a versatile integration platform that offers support for multiple integration paradigms. This is made possible by its ability to accommodate a range of interaction patterns and to provide a unified infrastructure for three major styles of Enterprise Integration - Service-oriented architectures, Message-driven architectures, and Event-driven architectures. In addition to this, ESB also provides advanced capabilities for mediating and transforming service messages and interactions, thereby enhancing the overall efficiency and effectiveness of the integration process.

2.3 WS02

WS02 ESB is an open-source enterprise service bus solution that allows for the integration of various applications and services. It supports multiple protocols and provides features such as message routing, transformation, mediation, security, and monitoring. It also has a graphical user interface for easy configuration and management of integration flows. WSO2's API management offering is available in two versions: on-premises (WSO2 API Manager) and public cloud (WSO2 API Cloud) [6].

2.4 Red Hat Fuse

Red Hat Fuse ESB is an open-source enterprise service bus solution that enables the integration of various applications and services. It provides a lightweight, modular, and flexible platform for building and deploying integration solutions. Red Hat has a strong presence in the application infrastructure market. It has now integrated a powerful API management platform with several application infrastructure offerings, including Red Hat Fuse (an integration platform), AMQ (a messaging platform), OpenShift (a container offering) and Red Hat Single Sign-On [6].

2.5 Fiorano

Fiorano ESB is a high-performance, scalable, and flexible enterprise service bus solution that enables the integration of various applications and services. Fiorano ESB supports multiple protocols and standards, including JMS, HTTP, SOAP, REST, and FTP. It also has a visual drag-and-drop interface for easy configuration and management. The Fiorano ESB implementation supports development in multiple programming languages. This, coupled with the inherent portability of the underlying infrastructure, makes the Fiorano ESB a true multi-language, multi-platform enterprise backbone.

2.6 Apache ServiceMix

Apache ServiceMix is an open-source enterprise service bus (ESB) and integration platform that provides a flexible and powerful infrastructure for integrating various systems, applications, and services. It is based on the Java Business Integration (JBI) standard and supports a wide range of protocols and technologies, including web services, messaging, and more. ServiceMix offers features such as routing, transformation, orchestration, and monitoring, making it suitable for building scalable and reliable integration solutions. It is widely used in enterprise environments for building and managing complex integration architectures. Apache ServiceMix is a flexible, open-source integration container that unifies the features and functionality of Apache ActiveMQ, Camel, CXF, and Karaf into a powerful runtime platform you can use to build your own integrations solutions. The software is distributed under the Apache License. ServiceMix supports the OSGi framework. ServiceMix integrated Spring Framework support and can be run at the edge of the network (inside a client or server), as a standalone ESB provider or as a service within another ESB. ServiceMix is compatible with Java Platform, Standard Edition or a Java Platform, Enterprise Edition application server. ServiceMix uses ActiveMQ to provide remoting, clustering, reliability and distributed failover. The basic frameworks used by ServiceMix are spring and XBean.

3. Comparative Analysis

3.1 Selection Criteria

The primary goal of this paper is to establish a framework for evaluating proposed ESB architectures or designs, in order to determine if they are capable of delivering the anticipated benefits within our organization and accordingly to compare and select the best possible ESB, to achieve this objective we will rely on the most widely accepted elements of the ESB definition as outlined in [8]. It is mentioned there are lot of criteria's that can be used but here defined some minimum or mandatory characteristics that should be exhibited by an Enterprise Service Bus.

3.2 Methodology

Out there many Enterprise service bus technologies (ESB) to choose from but out of all I have selected four ESB technologies to be compared. Based on their capabilities to be deployed in in-premise since our country legislation is not allowing cloud based services mainly in the financial sectors and also there is no ground to use cloud based ESB while all core services of the bank are on premise beside security related concerns when using open source technologies.

After short listing the ESB technologies I have selected four parameters for comparison among them and these are Ease of Learning, Features and Protocols, free resource & community support, and License & Cost. Parameters were explored through literature in the area of study, the contents found on web especially on particular technologies under our focus, looking on customer reviews who already deployed those technologies and by analyzing those customers' feedbacks and testing done on the technologies by setting up on small scale lab environments. Here it set parameters value both quantitative and qualitative means of measurement. Ease of Learning rated qualitatively as easy, moderate and steep considering the amount of time it takes to adopt the technology and the complexity nature of the system. Features and Protocols rated using high, few and average variable based on the number of features and protocols they can support with. Free resource and community support rated as low and high variables based on the availability of documents in the web and community pages. All parameters are combined together and presented in a tabular form.

License & Cost is crucial part for selecting final technologies and has been given the highest priority and technologies with highest priority will be chosen and recommended. The final ESB technology nominated shall improve the service integration and service orchestration at Tsehay Bank S.C

3.3 Learning Curve

When implementing new technology, it is important to study the learning curve of the artifact we are going to adopt. At present, teams working in core systems and Digital banking departments at Tsehay Bank has not yet applied features of ESB in full scale during application customization and integration. Sometimes even if the teams are conscious on the technologies and understand the rewards that can be exploited; it is worthless unless it is easily customizable, can be integrated without any hassle & complexity with minimum effort and time. It should be adopted within a very short period, with minimum and self-learning based approach. Otherwise, it impacts greatly the adoptability of the tool and discourages team's effort from quickly learning and applying the technology in their integration work. Hence, this need to be one of the main criteria when we select any artifact and familiarize ourselves with it.

The artifact selected here in this paper have comparable GUI and they look like to be

fairly easy to use. But, ESB, like all new technologies we encounter in our day to day operations, are always new unless we have prior knowledge and exposure.

In WSO2 the learning curve can be considered relatively good even if it requires understanding of concepts such as API management, message brokering, and service orchestration. However, WSO2 API Manager offers an intuitive and user-friendly interface that makes it easy to navigate and perform tasks. The layout and organization of features are well-designed, ensuring a smooth and efficient user experience. Whether it's creating APIs, managing subscriptions, or analyzing data, the interface provides a seamless experience for users. Apache ServiceMix is an open-source ESB that is based on Apache Karaf and provides a lightweight, flexible, and powerful integration platform. Its learning curve can be steep for those who are new to ESBs and related technologies, as it requires a good understanding of OSGi (Open Service Gateway initiative) and the Apache Karaf container. Overall, the learning curve for Apache ServiceMix can be significant for beginners.

The learning curve for Red Hat Fuse can also vary depending on the individual's background and experience with ESB technologies. For developers and IT professionals with prior experience in ESBs and integration platforms, the learning curve for Red Hat Fuse ESB may be relatively moderate. However, for those who are new to ESBs and integration technologies, the learning curve for Red Hat Fuse ESB may be steeper. And it demands them to familiarize themselves with the core concepts of ESB, such as message routing, transformation, mediation, and service orchestration, as well as learn how to configure and manage Red Hat Fuse ESB instances and integrate various systems and applications.

Fiorano has a learning curve that ranges from moderate to steeper depending on the individual, but its visual development tool makes it easier for non-technical users to understand to certain extent.

3.4 Features and Protocols

Each ESB tool discussed in the paper offers a rich set of features and supports various protocols with slight differences. WSO2 provides a wide range of features, such as API management, data integration, and support for microservices architecture. It also supports protocols like SOAP, REST, and WebSockets. Apache ServiceMix supports a wide range of protocols for communication and integration. Some of the key protocols supported by Apache ServiceMix include: HTTPS, JMS, SOAP SOAP, REST, FTP, SMTP, JDBC, AMQP and WebSockets. Fiorano stands out for its visual development tools, real-time messaging, and event processing capabilities, and supports protocols like JMS, AMQP, MQTT, and REST. Red Hat Fuse offers integration with other Red Hat products, containerization support, and cloud-native development features, while also supporting protocols like HTTP, JMS, AMQP, and MQTT. Each ESB tool has its own strengths and capabilities to cater to different integration needs.

3.5 Resources and Support Community

WSO2 offers comprehensive resource support with a dedicated team of experts who provide technical assistance, documentation, and training materials. They also have a large knowledge base, forums, and blogs for developers to find solutions to their queries. WSO2 also has a professional services team that offers consulting, implementation, and support services for enterprises.

WSO2 has a larger library of integrations compared to, which may be due to its longer presence in the market. The Apache ServiceMix community provides extensive support and documentation to help users understand, implement, and troubleshoot the platform through, mailing list, forums, issue tracker, documentation and community Events. Fiorano has a smaller community of users and contributors, but offers dedicated support services for enterprise customers. Red Hat Fuse benefits from the larger Red Hat community and support resources.

3.6 License & Cost

The tools discussed in this paper offer both open-source and enterprise options. For example, WSO2 has both open-source and enterprise versions, with pricing based on the number of production cores and support level. The Free License for Non-commercial, Educational, or Trial use is also available. Apache ServiceMix also free but there will be associated cost when professional support is need. The monthly average fee ranges from 200 to 500 USD for Red Hat Fuse and Fiorano ESB, which have relatively higher prices and depend on the number of cores and support level agreement.

Finally, all parameters are combined together in the table below.

Table 1: All Comparison parameters combined together

	WSO2	Apache ServiceMix	Red Hat Fuse	Fiorano
Learning Curve	Easy	Steep	Moderate	Steep
Features and Protocols	High	High	Average	Average
Resources and Support Community	High	High	Average	Low
License & Cost	Free	Free	\$500 per month	Free

4. Recommendation and conclusion

4.1 Recommendation

The digital banking division of Tsehay Bank S.C is dedicated to delivering cutting-edge services in the digital banking sector. In order to achieve this goal, the bank recognizes the need to reevaluate its use of ESB (Enterprise Service Bus) technologies or tools. As a new

bank, this presents an opportune moment to carefully select the right tool, minimizing the risk of future rework. Once the appropriate tool has been chosen and the setup process completed, subsequent integration efforts can be carried out without disrupting the established framework, thereby reducing costs and minimizing efforts. However, selecting the right tool is not a straightforward task, especially for a division with limited experience and exposure in the domain. To address this challenge, four prominent ESB options have been identified as potential candidates. These options have been compared based on criteria such as learning curve, features and protocols, resources and support community, as well as license and cost. Upon analysis, it is evident that WS02 stands out as the most favorable choice. With its user-friendly learning curve, extensive range of features and protocols, robust resource and support community, and cost-effectiveness due to its free licensing, WS02 emerges as the optimal choice to be selection based on the parameters defined.

4.2 Conclusion

The ESB tools discussed in this paper offer substantial benefits to Tsehay Bank's digital banking division, particularly in managing the increasing volume of service and application integration tasks. By leveraging ESB tools, the bank can save time, enhance operational efficiency, reduce service delivery timelines, and ultimately elevate customer satisfaction. Furthermore, utilizing ESB for integration work simplifies troubleshooting processes and facilitates adjustments to business rules and logic. Among the ESB options considered, I recommend WS02 due to its free licensing and robust ESB technology. While open-source products like WS02 may have limitations in terms of source code accessibility and potential vulnerabilities, especially in a corporate environment like Fintech, the availability of enterprise solutions provides a safety net when needed. It is prudent to initially familiarize the team with free options before considering enterprise solutions. However, it is important to note that this recommendation is based on theoretical analysis, as practical testing of the tool's performance was constrained by inadequate lab resources. Future work should include performance-related measurements as a significant parameter for the selection process.

References

- [1] Razvan Daniel , Radu Stefan , Laurenhu Ciovisa “Enterprise Service Bus-A Backbone for SOA”
- [2] Omer Aziz, Muhammad Shoaib, Adnan Abid , Naeem Asleam “ Research Trends in Enterprise Service Bus (ESB) Applications: A Systematic Mapping Study”
- [3] Afnan Alawami “Evolution of Enterprise Application Integration Technologies” (May 2022)
- [4] Periorellis, P., and P. Watson. "Enterprise Service Bus: An overview." School of Computing Science Technical Report Series (2007).
- [5] Wei, Ning. "Research on an information application integration platform based on SOA and web service." International Journal of Multimedia and Ubiquitous Engineering 10.12 (2015): 165-174.

- [6] Paolo Malinverno, Mark O'Neill, Aashish Gupta, Kimihiko Iijima. “ Magic Quadrant for Full Life Cycle API Management” (October 2019)
- [7] Fiorano Software Inc “Fiorano ESB Architecture” (2006 -2009)
- [8] Mark Endrei Jenny Ang Ali Arsanjani Sook Chua Philippe Comte Pål Krogdahl Min Luo Tony Newling “ Patterns: ServiceOriented Architecture and Web Services” (2004)
- [9] Adam Badura; Bartosz Sakowicz; Dariusz Makowski “Integration of management protocols based on Apache ServiceMix JBI platform” (2009 April)
- [10] Apache Service Mix <https://servicemix.apache.org/>
- [11] Omer Aziz, Adnan Abid , Naeem , Rubab Shaer “Trends in Enterprise Service Bus (ESB) Applications: A Systematic Mapping Study” (2020 February)
- [12] Tsehay Bank Website , <https://tsehaybank.com.et/company-profile/>

Distributed Database Selection for Vital Registration System of Addis Ababa

Mesfin Debebe

HiLCoE School of Computer Science & Technology
mesfindebebe@gmail.com

Abstract

This paper explores existing literatures on Distributed Databases and how they are implemented by different Database Management Systems. It will focus on identifying the basic properties of Distributed Databases and how they are addressed by two commonly used database management systems: Oracle and SQL Server.

The Vital Registration Agency of Addis Ababa has implemented a centralized database that is accessed by applications running at the different Woreda offices. The users at these offices are facing frequent problems in using the system due to network connection problems.

We propose that a distributed database is an ideal solution for such a problem and the aim of this research paper is to recommend the right technology to be used for the vital registration system in Addis Ababa. Hence, we will attempt to make a comparative analysis on the implementations of a distributed database using Oracle and SQL Server in order to recommend the better option based on the requirements and problems of the Vital Registration System of Addis Ababa.

Keywords: Distributed Databases; Distributed Database Management Systems; Oracle; SQL Server; Distributed Query; Distributed Transaction; Vital Registration System;

1 Introduction

The definition of a Distributed Database (DDB) as given by Özsu and Valduriez [1] is a collection of many databases that are connected by a computer network and a Distributed Database Management System (DDBMS) as a software system that manages these databases that are connected over the network. The different databases need to be logically interrelated and the DDBMS has the responsibility of making the distribution of data over multiple databases transparent to the user or applications. The DDBMS should also have the capability of using the processing and storage capacity of all the different database management systems that maybe located at different location

The different Database Management Systems in a DDBMS may be similar (homogeneous) or different (heterogeneous), but the DDBMS has to have the ability to manage them so that they appear as one database to the user. The databases are

interconnected by a computer network, and they cooperate in performing certain tasks. The Distributed Database Management System coordinates these tasks so that the distribution of the databases appears to be transparent to the user, i.e., the user is not aware of the existence of multiple databases.

Elmasri and Navathe [2] pointed out that a full-scale, dedicated and complete DDBMS that meets the techniques proposed in DDB researches are not yet been developed, and so they are not commercially available. Many Relational Database Management Systems (RDBMS) however, have tried to support the basic features of a DDBMS by including some functionalities to their products, e.g., Oracle and SQL Server.

Vital Registration Systems are systems where a government records the vital events of its citizens and residents. These systems are particularly involved in capturing and maintaining the basic information about each and every citizen and resident which may be residing in different parts of a country. The current Vital Registration System of the city of Addis Ababa uses a centralized database which is accessed from multiple application located at the Woreda sites.

The requirements of such a system can surely benefit from a DDBSs. The benefits of DDBSs becomes even more obvious when there are major problems of connectivity and the fact that data of residents is most likely to be processed at the location of their current residency.

We strongly believe that the existing problems of connectivity and availability can be greatly reduced by using a distributed database. We propose that a distributed database is an ideal solution for such a problem and the aim of this research paper is to recommend the right technology to be used for the vital registration system in Addis Ababa.

The remainder of this paper is divided into four sections. Section 2 briefly describes the general and specific objectives. Section 3 reviews relevant literatures related to Distributed Databases, their implementations and the Vital Registration system of Addis Ababa. Section 4 provides comparative analysis of the implementations of Distributed Databases using Oracle and SQL Server. Finally, section 5 will provide our conclusion.

Thus, the objective of this work is to examine problems of the current Vital Registration System of Addis Ababa and recommend a Distributed Database implementation by making a comparison between different implementations of a Distributed Database by different vendors, namely Oracle and SQL Server.

2 Related Work

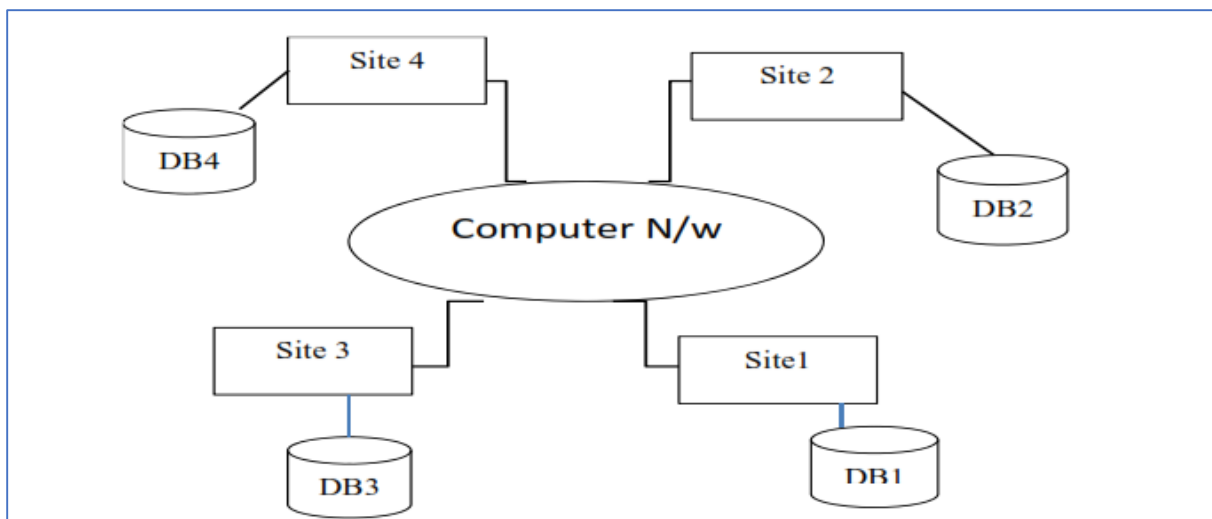
In this section selected research papers, reviews and articles on Distributed Database are briefly discussed. Focus will be made on the basic features of Distributed Database, different implementations of Distributed Database and the particular requirements of the Vital Registration System of Addis Ababa.

Although we have referred several literatures related to the Vital Registration System of Addis Ababa and Distributed Database and their implementations, the most relevant ones are summarized and cited.

2.1 Basic Features of Distributed Databases

Kaur and Singh [3] describe a distributed database management system (DDBMS) as a system that consists of a single logical database. The logical database is composed of a number of fragments and each fragment resides on its own database that is distributed over a computer network. Thus, we can think of a DDB technology as the result of a merger of two technologies: database technology, and network and data communication technology.

The figure given below, extracted from Tomar and Megha [4], illustrates the architecture of a distributed database management system.



In a Distributed Database, a user or an application that is located at one site is allowed to access data that is not found at the local site without actually knowing that the data is physically in another location. This is called distribution transparency. Another important aspect of a DDBMS mentioned by Grech [5], is that a user or an application that is located at one site will be able to process distributed queries; that is, a query should be able to reference more than one site.

We can say that one of the major advantages of a distributed computing systems is the partitioning of a big and unmanageable problem into smaller pieces and solve it efficiently in a coordinated manner. Elmasri and Navathe [2] site this as an additional factor to the economic feasibility of such approach. They claim that this advantage comes from two reasons: more computing power is used to solve a complex task, and each autonomous processing element can be managed independently.

Özsu and Valduriez [1] infer that the benefits of a DDBS and therefore the potentials of the technology can be summarized into four basic features. These are the transparent management of the data that is distributed over multiple databases, the consistent access to

the distributed data using distributed transactions, the improved performance that one gets from accessing local databases, and easier system expansion.

Elmasri and Navathe [2] summarize the basic features of Distributed Databases and conclude that they have to satisfy the following minimum conditions:

- There are different sites (computers) connected over a computer network.
- There are multiple databases and Database Management systems at the different sites
- The databases at different sites have to be logical interrelated.
- The database managements systems at the different sites communicate by sending data and queries (commands) over the network.
- The databases at the different sites may not be the same - in terms of data, hardware, and software.

Elmasri and Navathe [2] also outline some of the important reasons why organizations resort to distributed database management are the following [2]:

- Improved ease and flexibility of application development.
- Increased reliability and availability:
 - When a failure is encountered at one of the sites, other sites continue to operate.
 - Only the data and software that exist at the failed site cannot be accessed.
 - Therefore, in a distributed database, some of the data may be unreachable, but users may still be able to access other parts of the database.
- Improved performance:
 - A distributed DBMS fragments the database by keeping the data closer to where it is needed most. This reduces the delays encountered when accessing data in wide area networks.
 - Queries and transactions executed at a site have better performance not only because the data is not accessed over the network but also the because of the queries and transactions target a smaller local database.
- Easier expansion. In a distributed environment, expansion of the system is easier. Expansion could be in terms of
 - Adding more data
 - Increasing database sizes, or
 - Adding more sites is much easier

2.2 Implementation of Distributed Databases

Many Relational Database Management Systems (RDBMS) have tried to support the basic features of a DDBMS by including some functionalities to their products. This paper will review the implementation of DDBMS in Oracle and SQL Server.

Several DBMS solutions from various vendors exist on the market. In this paper we will compare and evaluate two of them that share major portion of the market: Oracle and Microsoft SQL Server. Oracle and Microsoft offer different versions and editions which can be used for different business needs. In this section, we will compare Oracle 11g and Microsoft SQL server 2005.

In order to make the evaluation of the products, we will start with the basic criteria for a distributed database system and filter these criteria using the particular case of the Vital Registration system of Addis Ababa.

Grech [5] states that the a distributes system provides the following transparencies [5]:

- Fragmentation Transparency: The user is not aware how the data is fragmented
- Location Transparency: The user is not aware of the physical location of the database
- Distributed Transaction Transparency: Transactions ensure the integrity and consistency of the database by concurrency and failure recovery
- Replication Transparency: The user is not aware of the replication of the data
- Support Heterogeneous System: Ability to connect with other databases which have different DBMS product
- Operating System: The Operating System on which the DBMS is capable of operating
- Cost: The cost of the solution

	Oracle	Microsoft SQL Server
Product Version / Edition	Oracle 11g Standard Edition	SQL Server 2005 Standard Edition
Fragmentation Transparency	NO	NO
Location Transparency	YES	Yes
Distributed Transaction Transparency	YES	YES (MSDTC)
Replication Transparency	YES	YES
Support Heterogeneous	YES	YES
Support Homogeneous	YES	YES
Operating System Support	Windows, UNIX, Linux	Windows

Functionality Comparison between Oracle and SQL Server

The above table shows that both Oracle 11g and Microsoft SQL server 2005 can be used

for the implementation of the Vital Registration system of Addis Ababa.

2.3 The Vital Registration System of Addis Ababa

The authors in [6] describe Vital Registration, also referred to as Civil Registration, as the process of recording the different important events about the residents of a nation as they happen. Some of the events that need to be recorded are the births, deaths, marriage, divorce, fetal death, annulment, judicial separation and adoption. The registration of these events makes them legitimate and accessible for statistical purposes.

Nielsen and Sahay [7] further enforce this concept and infer that the records resulting from Vital Registration systems have two major benefits. The first one is that they serve as personal legal documents that are required by each resident as proof of facts. Examples are a person's age, identity, children, etc. The second benefit is that the information derived from these records form the basis of a country's vital statistics system.

Hence, vital registration systems are used to derive the fundamental demographic and epidemiological measures that are needed in national planning across multiple sectors such as education, labor and health. They are also critical for a wide range of government activities (e.g., population registers and other administrative registers) and commercial enterprises (e.g., life insurance, marketing of products).

The World Health Organization (WHO) [8] has recognized a well-functioning Civil Registration and Vital Statistics (CRVS) information system as a key ingredient in strengthening CRVS systems in general. Further, the United Nations Commission on Information Accountability (COIA) for Women and Child health have especially mandated that countries would need to strengthen their CRVS information systems.

Furthermore, Setel, Macfarlane et al. [9] state that Civil Registration (CR) as being defined by the United Nations as the universal, continuous, permanent and compulsory recording of vital events. They also state that this registration process needs to be enforced through decree or regulation in accordance with the legal requirements of each country.

In the UN Resolution [10], the United Nations has adopted a resolution "Fundamental Principles of Official Statistics" on January 2014. This resolution states that member countries are required to keep official statistics. Principle 8 of this declaration states that "Coordination among statistical agencies within countries is essential to achieve consistency and efficiency in the statistical system".

Braa and Sahay [11] emphasis that the task managing Civil Registration and Vital Statistics systems is fundamental and that it requires a multisectoral approach. This should include at the minimum, the Ministries of Health, Justice and Finance. This approach calls for a collaboration between the different sectors and would help the different systems talk to each other and share data. With this regard, the World Health Organization [12] explicitly states that a holistic approach to the development of health information systems is required.

In line with this, the authors in [13] acknowledge that the Government of Ethiopia has given great emphasis to vital events registration across the country. Accordingly, it has put in place appropriate policies and allocated the required resources. Hence a system has been put in place up to the lowest administrative level. The effort included two Proclamations for the enforcement vital registration. The first is Proclamation No.760/2012: Registration of Vital Events and National Identity Card Proclamation. The second is Proclamation no.1049/2017: Vital Events Registration and National Identity Card proclamation (amendment) proclamation.

The Addis Ababa Vital Registration system was launched even before any legislative framework was formulated. This was done in 2008 under the shared responsibility of the Ministry of Science and Technology and the Vital Registration Agency of Addis Ababa. The responsibilities of the Ministry of Science and Technology has now moved to the Ministry of Innovation and Technology. Before 2008, there was no centralized database, and major problems in security and redundancy.

The authors in [6] indicated that the vital registration system of the city of Addis Ababa has put in place a centralized database where all the centers (woredas) of registrations are linked. They [6] further state that all these centers of registration along with the central database are connected over the woreda.net network. This woreda.net network is also being used by so many government agencies and hence is loaded with a lot of traffic.

The Woreda network is being administered by the Ministry of Innovation and Technology and the Vital Registration Agency has to report any network related problem of the system (which is frequent) to the Ministry of Innovation and Technology. This is also another major factor that contributes to the inefficiency of the system.

The existing system works with different type of data. Besides the textual, numeric and date information captured by the system, photographs and biometric identification like figure prints and iris scans are used. Currently, the system is being used for preparing the newly designed Digital ID. This new digital ID is now being printed at one location only, the central office.

3 Comparative Analysis

In this section a comparative analysis of the implementations of a Distributed Database using Oracle and SQL Server for the Vital Registration System of Addis Ababa will be made.

Both Microsoft SQL Server and Oracle Database are popular relational database management systems but can also be used for the implementation of distributed databases. Hence the comparison will be based on how well suited these technologies will be for the Vital Registration system for the city of Addis Ababa.

We will explore the basic features of Distributed Databases and select the ones that are more relevant to the Vital Registration system of Addis Ababa. Hence, our comparison will

be based on this set of parameters.

Based on the requirements and problems of the Vital Registration system of Addis Ababa, the following parameters have been selected to make the selection of a DDBMS for its implementation:

- Integration of the system with other systems is fundamental.
- This implementation should handle huge amount of data over the network.
- We should be able to send and retrieve data and commands among different Woredas/Kebeles easily and quickly.

3.1 Design Considerations of a Distributed Database

There are different factors that need to be considered when designing a system. The design has to take into account the particular need of the system. The same is true in designing Distributed Database system. We should not only consider how the system is to be distributed, but also how the specific issues of the problems are to be addressed.

We have identified the following main features as major factors in the design of the Distributed Database system for the Vital Registration System of Addis Ababa. These are: Connectivity, Fragmentation, Distributed Transaction and Replication.

Connectivity: The Distributed Database system needs to connect to both homogenous and heterogeneous databases.

Fragmentation: Fragmentation is the partitioning of the database data structures into smaller portions. Three types of fragmentation exist: horizontal fragmentation, vertical fragmentation and mixed fragmentation. In the case of the Vital Registration System of Addis Ababa, Horizontal fragmentation will be used to allocate data into their respective sites (Woreda/Kebele). Meaning that the data for each resident will be stored according to the place of residence i.e., it will be located at the sites where it would be mostly used.

Distributed Transaction: It is believed that when data is stored horizontal fragmentation according to the place of residence, it will significantly reduce the concurrency problem that may be encountered in a distributed database implementation. Thus, the system can use serializability which means that transactions are executed serially, without any major conflict.

Replication: Replication is the process of maintaining multiple copies of data items at different locations. When changes occur, they are captured at the local database. These changes are then sent to the other databases or to the main database at a convenient time. Hence, since several copies exist, if a site becomes unavailable, the DDS can still respond to queries using the replicas.

Though replication is not a main feature of a Distributed Database system, it is a major consideration in its design in order to provide high availability and fault-tolerance.

Support to Standard Operating Systems: It is preferable for the Distributed Database Management System to support the Standard Operating Systems.

According to the design consideration stated in the previous section Connectivity, Fragmentation, Distributed Transaction, Replication and Support to Standard Operating Systems were identified as the critical factors for the implementation of the Vital Registration System of Addis Ababa.

Connectivity with homogeneous database systems is supported by both Oracle and SQL Server. Oracle makes use of database links while SQL Server uses linked server links to connect with other databases. In case of heterogeneous database connectivity, generic connectivity such as ODBC is provided by both Oracle and SQL Server. Oracle provides a larger number of possible data sources when compared to SQL Server. The Vital Registration system may involve only homogeneous databases, but it is anticipated that integration with other systems may be required in the future. Both Oracle and SQL Server provide connectivity to heterogeneous systems but Oracle provides a larger number of possible connections.

Though both Oracle and SQL Server do not have built-in features to distribute the fragments, the distribution can be included in the application logic. But they still provide location transparency which can be achieved by the use of synonyms and views in both Oracle and SQL Server.

Distributed Transactions are supported by both solutions by using the two phase-commit mechanisms. In the case of SQL Server, an additional component (Microsoft distributed Transaction Coordinator - MS DTC) has to be installed and configured.

Oracle uses its Oracle Streams and materialized views to support Replication while SQL Server provides three types of replications which are Transactional, Merge and Snapshot replication.

Kaur and Singh [3] report that Oracle supports Windows, UNIX and Linux Operating Systems while SQL Server support only the Windows Operating System. Microsoft's MSDN [14] claims that, beginning with SQL Server 2017, the platform is now available on Linux as well.

Another factor that is considered in the implementation of databases is the size of the data that needs to be stored. In this case the size of the data that need to be captured for each resident and the total population of the city needs to be considered.

In regards to the size of the data that need to be captured for each resident; we have made an estimation based on the data structure of the current system. The authors in [13] state that the basic registrations are done on the events of birth, mirage and death. Most of the data capture during these events are alpha-numeric and do not have an extensive need of massive storage capacity.

The National Central Statistical Agency [15] reports that though there are different estimations of the population of Addis Ababa, some taking it as high as close to 5 million. However, based on the last census on 2015, the population of Addis Ababa is estimated to be 3,273, 000 (CSA, 2015). Furthermore, this population is distributed into more than 10 Sub-cities.

With this amount of population and the data that needs to be captured for each individual, the total amount of capacity required for the databases at each Sub-city or Woredas will not be an issue for both Oracle and SQL Server.

Another factor is the price of the products. The Enterprise edition of Oracle costs about 20,000USD per year while Enterprise edition of SQL Server costs about 12,000USD per year [16], Definitely, SQL Server is significantly cheaper than Oracle.

4 Conclusion

Based on our findings, both Oracle and SQL Server have their own features for implementing a Distributed Database system for the implementation of the Vital Registration System of Addis Ababa. We can say both provide a feasible solution, even when we consider the particular needs and problems of the Vital Registration System of Addis Ababa. The ultimate breaking point boils down to cost, and SQL Server is significantly cheaper than Oracle.

Thus, our recommendation for the choice of a Database Management system for implementing a Distributed Database system for the Vital Registration System of Addis Ababa is Microsoft's SQL Server.

It should be noted that our recommendation is solely based on the features of Oracle and SQL Server in relation to the requirements of the Vital Registration System of Addis Ababa, Recent versions of Oracle and SQL Server may also have improved or new feature that complement the implementation of Distributed database systems.

References

- [1] M. Tamer Özsu and Patrick Valduriez, *Principles of Distributed Database Systems*, Springer Science & Business Media, LLC 2011.
- [2] Ramez Elmasri & Shamkant B. Navathe, *FUNDAMENTALS OF Database Systems*, 6th Edition, 2011.
- [3] Kirandeep Kaur and Harwant Singh, "Distributed database system on web server: A Review" *International Journal of Computer Techniques* — Volume 3, Issue 6, Nov - Dec 2016.
- [4] Parul Tomar and Megha, "An Overview of Distributed Databases" Volume 4, Number 2, 2014.
- [5] Joseph Grech, "Designing and Implementing a Distributed Database for a Small Multi-Outlet Business", Regis University, Summer 2009, Available at: <https://epublications.regis.edu/theses/5>

- [6] “Rapid assessment of national civil registration and vital statistics systems”, Technical document, 9 May 2010, Available at: <https://www.who.int/publications/i/item/rapid-assessment-of-national-civil-registration-and-vital-statistics-systems>
- [7] Peter Nielsen and Sundeep Sahay , “CIVIL REGISTRATION AND VITAL STATISTICS: A NEGLECTED AREA IN INFORMATION SYSTEMS RESEARCH”, University of Oslo, Norway, Murodillo Latifov, EPOS Health Management, Tajikistan, Proceedings of the 13th International Conference on Social Implications of Computers in Developing Countries, Negombo, Sri Lanka, May 2015
- [8] World Health Organization WHO, “Keeping promises, measuring results: Commission on information and accountability for Women’s and Children’s Health” , Report, Geneva, 2011
- [9] Setel, P.W., Macfarlane, S.B., Szreter, S., Mikkelsen, L., Jha, P., Stout, S., and AbouZahr, C. 2007. “A scandal of invisibility: Making everyone count by counting everyone”, *The Lancet*. 370, 9598, 1569-1577.
- [10] United Nations, General Assembly Resolution, “Fundamental Principles of Official Statistics”, A/RES/68/261, 73rd plenary meeting, 29 January 2014
- [11] Braa, J. and Sahay, S, “Integrated health information architecture: Power to the users”, Matrix Publishers, Delhi, India, 2012
- [12] World Health Organization, “World health statistics 2007”, Geneva, 2007.
- [13] UNICEF Website, Available at: <https://data.unicef.org/crvs/ethiopia/>
- [14] Microsoft MSDN, <https://docs.microsoft.com>
- [15] National Central Statistical Agency, (2015), Available at: <https://www.statsethiopia.gov.et>
- [16] Oracle Database vs. Microsoft SQL Server, Available at: <https://www.datamation.com/>

Communication from the Editorial Board of HJCST

Dear Esteemed Researchers and Members of the Scientific Community,

The Editorial Board of the *HiLCoE Journal of Computer Science and Technology (HJCST)* is pleased to reach out with an important message concerning the dynamic landscape of emerging technologies and the critical role of research in shaping the future.

In recent years, rapid advancements in **Artificial Intelligence (AI)**, **Big Data Analytics**, **Cybersecurity**, **Cloud Computing**, **Blockchain**, **Internet of Things (IoT)**, and **Quantum Computing** have begun to transform industries and redefine the frontiers of computer science and technology. These domains are not only influencing how we build intelligent systems but are also presenting new challenges and opportunities for innovation.

We would like to highlight the following key areas that are of growing relevance and societal impact:

- **Artificial Intelligence & Machine Learning:** From deep learning and natural language processing to responsible and ethical AI, this domain continues to be at the forefront of technological advancement.
- **Big Data & Data Science:** With the explosion of data, novel methods of data mining, real-time analytics, and predictive modeling are crucial for decision-making across sectors.
- **Cybersecurity & Privacy:** As digital threats grow more sophisticated, research into secure architectures, cryptographic methods, and privacy-preserving systems is essential.
- **Cloud Computing & Edge Computing:** Innovative resource management, scalability, and latency-reducing techniques are key to supporting modern applications and services.
- **Blockchain Technologies:** Beyond cryptocurrencies, blockchain presents opportunities for secure, decentralized systems in areas like supply chain, healthcare, and digital identity.
- **Internet of Things (IoT):** The integration of devices and sensors into smart systems demands research into connectivity, scalability, and security.
- **Quantum Computing:** Although still in its early stages, quantum computing promises breakthroughs in computational power, with implications for cryptography, optimization, and more.

It is imperative that the **Ethiopian research community** actively participates in these global technological trends by engaging in **applied research** that addresses local challenges, leverages indigenous knowledge, and contributes practical solutions to national development goals.

As one of Ethiopia's **leading private institutions**, *HiLCoE School of Computer Science and Technology* is dedicated to becoming the country's premier center of **IT education, research, and consulting**. In line with this vision, the School strongly encourages researchers, academics, and practitioners to take part in shaping the future through innovative and high-impact research contributions.

The *HiLCoE Journal of Computer Science and Technology (HJCST)* is committed to advancing knowledge and fostering innovation. We **invite you to submit scientific papers** in the aforementioned areas for consideration in **future issues of the journal**. Interdisciplinary research, novel applications, theoretical breakthroughs, and case studies with practical relevance are all welcomed.

By engaging deeply in these evolving fields, we can collectively contribute to shaping a more intelligent, secure, and data-driven future.

We look forward to your valuable contributions.

Sincerely,
Editorial Board

HiLCoE Journal of Computer Science and Technology (HJCST)