A Framework for an E-Health System for Zambian Health Centres that incorporates Data Mining Reporting

Selina Kadakwiza Halubanza[1], Douglas Kunda[2] and Brian Halubanza[3]

1,3 Department of Computer Sciences, School of Engineering and Technology, Mulungushi University, Kabwe.
2 Department of Computer Science, ZCAS University, Lusaka, Zambia,
Corresponding Author Email: selina.halubanza@gmail.com

Abstract: ***The implementation of e-health systems enables healthcare organizations to streamline multiple processes and deliver services more efficiently, resulting in cost savings. Health facilities in Zambia have adopted SmartCare e-health to manage the out-patients but there have been some identified technical and organisational challenges. The current SmartCare is not web based hence making it difficult for health personnel to access it anywhere since the card readers are only stationed in health facilities. The cards are susceptible to being destroyed as they are inserted in various card readers. SmartCare has no centralised database which could also provide real time patient data. This study sought to create a web-based e-health system for a health centre in Zambia, which includes reporting through data mining. This is achieved by use of a centralised database that can be accessed anywhere at any time. In addition, the proposed design integrates python engine to facilitate data mining reporting. Design Science and Data Mining Model development research methodology was adopted for this study.***

Keywords: E-health, SmartCare, Data mining, Data mining tools and techniques, Web application, Zambia

1. Introduction

E-health is expanding in the fields of medical informatics, public health, and internet-based health services, driving global advancements in technology to address long-standing issues, reduce expenses, and enhance patient care. For instance, the use of IoT-cloud-based e-health systems with underlying IoT networks enables communication among users, services, and servers, with medical data stored in the cloud. This type of system enhances healthcare services and encourages ongoing systematic innovation. [1].

Health organizations have many of their processes simplified because of e-health services providing a more efficient and cost-effective manner [2]. E-health improves the accessibility of clinical data for health professionals across all levels, enabling informed decision-making. It also promotes the transparency of information as a means of feedback for ongoing health enhancements within communities, among individuals, and throughout the healthcare system[3].Electronic Health Records (EHR) has improved the quality of care due to reduced medical errors and it has also provided healthcare workers with decision support [4].

In partnership with the Centres for Disease Control and Prevention (CDC) and various implementing partners, the Ministry of Health (MOH) in Zambia developed and implemented the SmartCare EHR System. This system is a comprehensive electronic health record that offers seamless care and a clinical management information system at the facility and district level. It serves as a key element in a national monitoring and evaluation framework. Presently, SmartCare is operational in approximately 963 facilities in all districts of Zambia. While the partners are aiding its deployment in both government and private facilities, the government deployments and enrolment rates are witnessing the fastest growth[5].

Unlike most systems that have a centralized design, SmartCare data at each facility is stored in a distributed design, which means that an Internet connection is not mandatory for its usage. The system employs client care cards, which contain an individual's health information, to facilitate the maintenance of care continuity across visits, health facilities, and health services. As an "e-first" system, SmartCare relies on computer-based data entry by health staff for capturing medical information [6]. Each patient is provided with a SmartCare card that contains an embedded chip carrying all their health information, along with a copy of the information from the clinic [7].

The SmartCare electronic system was developed to enhance continuity of care and provide timely data on maternal and child health, HIV/AIDS, tuberculosis, and malaria interventions for public health purposes. Patient details are captured and retrieved electronically from the SmartCare system, enabling authorized personnel to easily access patient records and expedite medical staff attendance to patients. When patients visit a healthcare facility that utilizes the SmartCare system, they are requested to register for a healthcare card that contains all of their patient data. However, the current SmartCare system is not web-based, thereby limiting healthcare personnel's access to it from anywhere outside of healthcare facilities. Additionally, the physical SmartCare cards are prone to damage as they are frequently inserted into various card readers.

SmartCare lacks a centralized database capable of providing real-time patient data, and its current version lacks a data mining reporting feature. Moreover, specialized statistical analysis and data mining reporting require extracting data from SmartCare. Hence, the objective of this study was to develop a web-based e-health system for Zambian health centres that incorporates data mining reporting. The study aimed to achieve the following research objectives: (a)

design a web-based e-health system for Zambian health centres, (b) evaluate critical data mining tools and techniques for e-health in Zambia, and (c) integrate important data mining tools and techniques into the web-based e-health system for Zambia.

2. Literature review

2.1 Health system

The quality of health information generated is closely linked to the performance of healthcare systems. High-quality health information enables health systems to make informed decisions, plan effectively, monitor and evaluate progress, conduct epidemiological surveillance, and more. Every year, the World Health Organization (WHO) publishes global health statistics to measure health outcomes in member countries. Therefore, it is crucial to have a well-structured Health Information System (HIS) that can adequately produce, analyse, store, and share reliable and accurate information for decision support across all levels of the healthcare system [8].

At health centres, patient records are typically stored in cabinets. When a returning patient visits the centre, their file is retrieved from the cabinets, and any new information is added to the existing documents. If the file cannot be located, a new file is created for the patient [9].

2.2 E-health systems

In partnership with the United States Agency for International Development (USAID), the Ministries of Health and Community Development, Mother and Child Health in Zambia, and the Ministry of Health and Social Welfare in Tanzania developed an open-source Electronic Logistics Management Information System (eLMIS) [10]. The eLMIS is an online application that is accessed through a web browser, and its data is stored centrally in a database. Through eLMIS, health facilities are connected to the central store, enabling real-time collection and distribution of logistics data. This system helps supply chain managers identify which medicines are being used and what is required, thereby ensuring a continuous supply of medicines to patients. By transitioning from a paper-based system of data management to an electronic format, eLMIS has facilitated better, faster, and more accurate reporting of supply chain data, reduced stock outs of health commodities, and ultimately improved access to medicines, thereby enhancing health outcomes [11].

The rapid expansion of the Internet of Things (IoT) has had a significant impact on our daily lives, especially in the field of electronic health (e-health). Patients with chronic or severe illnesses are often fitted with implants or medical sensors to monitor different types of physiological data. These medical devices are connected to the e-health IoT network, and the collected physiological data are transmitted to the patient's gateway device over the internet, which forms the patient's electronic health record (EHR)[12]. As the volume of data in electronic health records (EHRs) of patients increases, challenges arise in terms of information privacy, search, updating and sharing. To address these challenges, a privacy-preserving e-health system was designed that integrates the Internet of Things (IoT) and big data. The system assigns anonymous identities to both patients and medical nodes, which are calculated from their real identities. In case an anonymous patient is found dishonest or misbehaving, the trusted authority can trace their real identity. Similarly, if an anonymous medical node is used to launch an attack in the patient's IoT network, the patient can recover the node's real identity. The confidentiality of messages transmitted in the health IoT network is guaranteed by the patient generating a symmetric key and sending it to all the medical nodes in a privacy-preserving way. The e-health big data are encrypted and stored in a cloud platform, and the system has an expressive and lightweight fine-grained access control mechanism to prevent unauthorized data access. The patient controls the EHR encryption procedure and defines an access policy such that only data users with specific attributes can decrypt their medical files [12]. While big data analytics presents significant opportunities for the healthcare sector to leverage data-driven solutions to tackle healthcare challenges, a study conducted by [13] revealed that several healthcare providers have not yet fully utilized these systems due to factors that hinder the implementation of big data analytics. Technological inadequacy, personnel shortages, data management issues, policy limitations, power outages, low employment levels, and challenges in replacing faulty parts are some of the factors that affect the implementation of big data, as per the study conducted on the Copperbelt province of Zambia.[14] A prototype was developed to enhance the provision of healthcare services, which included tasks such as scheduling appointments, processing payments, storing medical records, providing information, and processing orders for medicine products. To ensure patient privacy, a security plan was implemented to address emerging threats and prevent unauthorized access to medical records through encryption and access controls [15] proposed a model for an electronic health management information system that supports structural interoperability in heterogeneous environments to facilitate continuity of care. The model proposes a service-oriented architecture that uses RESTful web services with JSON to achieve scalability, low cost, interoperability, and high availability for eHealth systems.

2.3 E-Health in Zambia

The Zambian health service delivery system operates across three levels [16]. The first level is the primary health care services which are located within communities and districts and serve as the initial point of contact for patients. The second level comprises provincial level facilities which function as referral points from the primary level. The third level includes teaching and specialized referral hospitals.

The Ministry of Health in Zambia recognized the potential benefits of e-Health and developed a strategy to explore effective ways to respond to the healthcare needs of individuals and communities. The implementation of e-Health has improved the delivery of health services by overcoming barriers resulting from inadequate personnel, geographical constraints, and physical inaccessibility to health facilities. It provides the necessary infrastructure for

information exchange between participants in the healthcare system and serves as a driver for improved health outcomes. [5].

The e-Health strategy gives clear guidelines to the ministry that drives growth and change over the effective use of ICTs. The strategy gives wide-ranging operational guidelines on how to make the various e-Health systems efficient and interoperable. The priority focus areas include Service Delivery, Research, and eLearning [5].

1. As reported in [16], the current national health management system in Zambia collects data only from the health-facility level and above, as well as some data from health posts reported by Community Health Assistants (CHAs). Unfortunately, the data collected by Community Health Workers (CHWs) is not captured by the system. Although a paper-based reporting system exists for transmitting data from CHWs to health facilities, it is not fully functional due to challenges such as sustaining the availability of reporting forms and delivering completed forms to the health centres. To address this issue, Zambia employs a double level system that involves monthly information gathering from the community and weekly information gathering from the facility level, using a combination of paper-based forms for data aggregation before transmitting data using inexpensive java-based feature phones. The Health Management Information System (HMIS) is used to process the data collected from medical records, whether they are paper-based or electronic, and produce statistical reports [17].

2.4 What is SmartCare?

In partnership with the Centres for Disease Control and Prevention (CDC) and other implementing partners, the Ministry of Health (MOH) in Zambia created and implemented SmartCare, an Electronic Health Record System (EHR). SmartCare is a comprehensive EHR system that ensures continuity of care and offers a clinical management information system at both facility and district levels. It serves as a crucial component in the "one National M&E system" [5].

SmartCare is an electronic health record system that is designed to be an ‘e-first’ system, meaning that health staff use computers to input medical information. Patients are provided with a SmartCare card that has a chip which stores all their health information, along with a copy of the information from the health facility [7]. The development of the SmartCare software in Ethiopia was a collaboration between the SmartCare teams in Zambia and the United States. Ethiopia adopted SmartCare in 2007 with the support of Tulane University Technical Assistance Project Ethiopia (TUTAPE). SmartCare is capable of supporting longitudinal record-keeping for various health issues such as HIV/AIDS treatments, TB care, VCT, and antenatal care. The system also offers clinical decision support, touchscreen interaction, offline data synchronization, and data portability through the use of smart cards [18].

The healthcare sector has recognized the significance of e-Health as an integral component, owing to the growth of the Internet and advancements in networking and information communication technologies. According to the World Health Organization (WHO), e-Health is the utilization of information and communication technology (ICT) to connect patients, providers, and governments. It aims to educate and inform healthcare professionals, managers, and consumers, promote innovation in care delivery and health system management, and enhance our healthcare system [19].

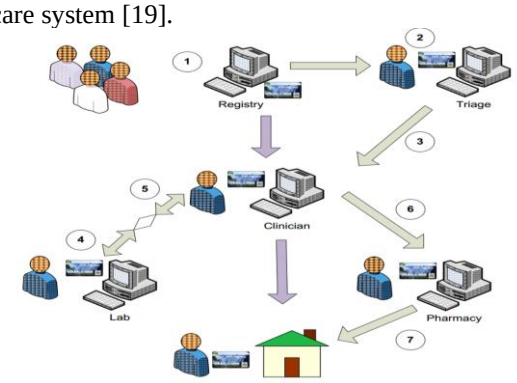

Fig 1. Client comes for first time

The list of the activities when a patient comes for the first time at health facility (see Fig 1) includes:

1. Client is registered from first point of service (e.g., Registry) where care card is issued.
2. Client moves to next point of service (e.g., Triage) and presents care card to provider.
3. Client Moves to next service point and presents card to clinician who inserts card (authenticates user) and enters visit details.
4. Client proceeds to lab Presents care card to lab tech Lab inserts card (authenticates user) view lab order Collect lab sample Updates client card with lab results.
5. Client returns to clinician presents card Clinician views results and continues service.
6. Client goes to pharmacy Presents care card to pharmacist inserts card (authenticates user) view prescription Pharmacist dispenses drug.
7. Client goes home with up to date care card.

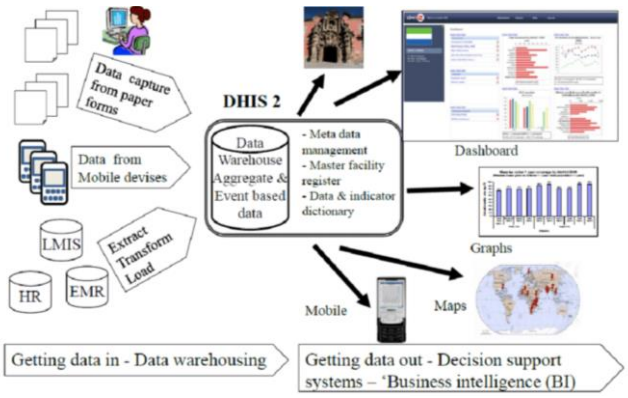

Fig 2. DHIS2 Architecture [20]

2.5 District Health Information System 2 (DHIS2)

DHIS2, a web-based open-source platform, is utilized by 67 low and middle-income countries (LMIC), non-profit organizations, non-governmental organizations (NGOs), and multinational organizations for health management information system (HMIS). The architecture of DHIS2 is depicted in Figure 2. DHIS2 has evolved from being a platform that collects, stores, validates, analyses, and presents aggregated health data to supporting patient management, monitoring and evaluating health services, and assessing the health status of populations, according to [21]. DHIS2 serves as a foundation for LMICs to achieve Sustainable Development Goals, and it has become the primary source of quality data for public health resource allocation and impact measurement [22]. However, challenges and concerns persist, as data is typically gathered at health facilities by overburdened healthcare workers and later manually tallied and entered, resulting in delays and inaccuracies [22]. This issue is also prevalent in most health centres in Zambia.

The Zambian health management information system is based on the DHIS2 database platform, which integrates various data capture modes, aggregated forms, entity tracking, and single event data, as well as efficient data warehousing and analytics tools [16]. DHIS2 is designed to serve as a standardized collection and analysis tool that is widely used across diverse health programs to support decision-making at all levels of health services. Despite the challenges posed by differences in resources and organization, DHIS2 is aimed at bringing together all routine reporting onto a single platform [23]. DHIS2 has introduced several features, including dashboards, instant visualization tools, GIS, and pivot tables, to enhance the utilization of data for decision-making purposes [20].

2.6 Data Mining

Data mining has become an integral part of the healthcare sector, allowing important data to be extracted and analysed. This analysis can help identify prevalent diseases, their causes, symptoms, precautions, and remedies, which can aid in preventing or minimizing their impact. By applying data mining techniques, healthcare management systems can better identify high-risk diseases and design appropriate interventions, as well as track chronic disease states. Data mining can also reduce hospital admissions and claims, shorten patient stays, improve medical practices, and enhance patient outcomes, all in a cost-effective manner. For instance, a multinomial Naïve Bayes algorithm was proposed to detect heart failure using a data set of 30 variables, and compared against several classification algorithms, such as Neural Network, Logistic Regression, Random Forest, Decision Tree, and SVM. These techniques can help healthcare practitioners make intelligent clinical decisions that surpass traditional decision support systems, resulting in improved patient care.

In the healthcare sector, data mining techniques can address complex queries related to diagnosing heart disease, and the results of data mining can help healthcare practitioners make intelligent clinical decisions, which may be more effective than traditional decision support systems [26]. Additionally, various data mining techniques such as Navie Bayes, MLP, Bayesian Network, C4.5, Amalgam KNN, ANFIS, PLS-LDA, Homegenity Based, ANN, and Modified J48 have been used to explore early prediction of diabetes [27].

2.7 Data Mining Techniques and Tools

Data mining techniques are useful for analysing large amounts of health data to uncover significant patterns and predict outcomes using both descriptive and predictive models [28]. Predictive modelling involves identifying patterns within the data to forecast future values. Various types of models such as classification, regression, and AI-based models fall under predictive modelling. These models are built by training them with data where the response variable's value is already known. Descriptive models, on the other hand, belong to the realm of unsupervised learning. They explore the database to discover patterns and relationships in the data. Clustering (segmentation) algorithms, pattern recognition models, visualization methods, and others belong to this category of descriptive models. By using complex algorithms, data mining techniques can explore, analyse, and extract useful medical data to uncover unknown patterns [29]. Data mining techniques can also be applied to analyse various factors that contribute to diseases, such as the type of food, work environment, education level, living conditions, and access to clean water, availability of healthcare services, and cultural, environmental, and agricultural factors [30]. KNIME, R Project, RapidMiner, Scikit-learn, and Spark are among the most commonly used open-source data mining tools, as reported by [28]. This research integrated classification, association, disease prevalence, and incidence into a web-based e-health application.

2.7.1 Classification

Classification is a data mining technique that involves learning from a set of cases to make accurate predictions of the target class for new cases. This process involves two steps: (1) the training (or learning) phase and (2) the test (or evaluation) phase, where the predicted class is compared with the actual class of the instance. Although classification is usually performed using supervised learning, it can also be carried out through unsupervised learning, such as in the case of clustering techniques where the class is unknown or unused [31].

2.7.2 Association

Association rule mining is a technique that can effectively identify relationships between different items, and it has been applied successfully in healthcare [32]. Several statistical algorithms have been developed to implement association rule mining, with Apriori being one such algorithm. As stated in [33], association rule mining aims to uncover relationships among many items in the database and identify the interrelationships between all frequent items or attribute subsets. In a study, the researcher proposed the use of association rules in the application of disease complications. Through association rule mining, the goal was to define the relationship between healthy data and disease, the relationship between the disease and its potential complications, and to identify frequent relationships between different complications of the disease. Ultimately, this would provide a more comprehensive decision support system for wise medical decisions.

The Apriori algorithm is a technique that can detect and examine the underlying patterns of different sets in a database. It is typically divided into three categories: simple association, temporal association, and causal association [12]. The Apriori algorithm works by first extracting information from a database, then dividing it into frequent items and itemsets, and generating candidate sets for association rule mining. This process is carried out with the aim of obtaining the minimum value of support and minimum confidence value [34].

The FP-Growth algorithm is an improvement over the Apriori algorithm in terms of execution speed, which helps to overcome the limitations of the Apriori algorithm. Frequent Pattern Growth (FP-Growth) is an alternative algorithm used to identify the most frequently occurring set of data (frequent itemset) in a dataset. FP-Growth is used as one of the algorithms to solve Association Rule problems. The algorithm consists of two stages: first, compression is performed on the database based on frequently occurring items to create a Frequent Pattern Tree (FP-Tree). Second, separation is performed on the database results to obtain a compressed database conditionals [35].

2.8 Prevalence and Incidence

Prevalence refers to the proportion of a population that has a disease at a specific point in time. It is calculated by dividing the existing cases of a disease at a certain point in time by the population that is at risk of having the disease. Prevalence takes into account both new and existing cases of the disease [36].

On the other hand, incidence refers to the occurrence of new cases of a disease in a population during a specified period. Incidence only considers new cases of the disease. It is calculated by dividing the new cases of a disease in a defined population over a period (such as a day) by the population at risk [36].

2.9 Application of data mining in e-health

In a study reported by [37], data mining techniques such as classification and clustering were utilized to detect diabetes, digestion, and kidney diseases in a dataset. The study also compared the results of these techniques with those obtained using the decision tree technique for the same dataset. Another study conducted by [38] employed classification techniques to analyse Hepatitis-C patients and predict their survival rate. The study aimed to determine whether patients with hepatitis are likely to survive or not. Furthermore, classification techniques were also used to distinguish between Dengue and Chikungunya patients.

The reduction of maternal, neonatal, and infant mortality rates is a critical objective in health data monitoring, as stated in [39]. The Brazilian health databases contain significant data that can be used to predict the risk of death during the early stages of gestation and infant development. The researchers developed various death risk prediction models based on the availability of information during the gestational period, aimed at the public of interest. They used machine learning techniques to classify death risk for maternal, neonatal, and infant patients. They also presented an experiment pipeline to evaluate machine learning models using different feature combinations to estimate the average performance. The study showed that Random Forest performed better than other machine learning methods [39].

In [40], various machine learning techniques were utilized to detect three different diseases: Dengue, Diabetes, and Thyroid. The classifiers used in the study included Decision Tree, Gaussian Naive-Bayes, Random Forest, Logistic Regression, k-Nearest Neighbors, Multilayer Perceptron, and Support Vector Machine.

3. Methodology

The research followed the methodology of developing a Data Mining Model and Design Science [41]. This approach was chosen as the goal of the study was to create and implement an artifact, which in this case was a web application. The Design Science research paradigm involves a designer addressing human problems by creating new and innovative artifacts, thus contributing to the advancement of scientific knowledge. The created artifacts are considered essential in comprehending the problem at hand and are equally valuable in practical applications.

3.1 Design Science Research Approach (DSR)

The gap between relevance and rigour in information systems research can be effectively addressed by DSR, or Design Science Research [42]. DSR is a research paradigm that aims to create innovative artifacts and processes to improve the environment. A good DSR starts by identifying opportunities and problems in a real application environment [41], which is the approach adopted in this study. The objective of this research was to design an E-Health System for Health Centres that incorporates Data Mining Reporting. The six phases of DSR proposed by [43], including problem identification, solution objective definition, design and development, demonstration, evaluation, and communication, were followed in this study. The research design framework adopted in this study is presented in Fig 3. DSR is increasingly becoming popular and finding its application in information systems research.

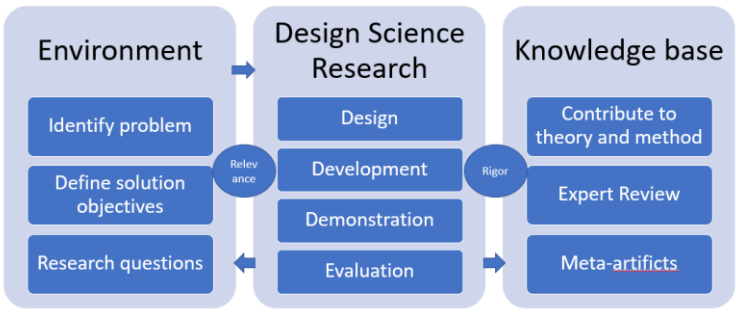

Fig 3: Research Design Framework
(Adapted from [43])

3.1.1 Identify Problem,

As noted from Fig 3 above, the first two processes encompass the proposal in which the objectives of the study to the problem statement is identified and that this activity was done under problem statement.

3.1.2 Define the Solution Objectives

The solution objective and requirements were defined. The following design artifact requirements were elicited administration and staff can login into the system. Staff can add patient, edit, and view e-health information (patient details, outpatient details, inpatient details, xray details, record of birth details, Hiv details, death record details, medicine details, staff details, covidtesting details, ward details, treatment details, observation details, adultpaediatricart, antenatal details, family planning details, underfive details). Staff can view patient information, staff statistics, diseases statistics, e-health reports, and data mining reports.

3.1.3 Design and Development

A web-based e-health system was designed using a use case diagram (refer to Fig 4). This diagram illustrates the relationships that exist among actors and use cases within the system. The use case diagram provides an overview of the usage requirements for a system or organization, which can be in the form of an essential model or a business model. It is a useful tool for identifying the scope of a project development and modelling the analysis of the usage requirements of a system, as documented by [44].

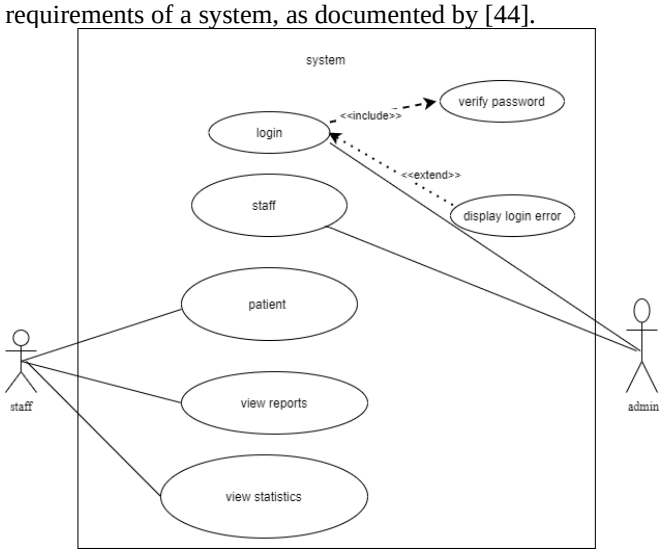

Fig 4 Use Case Diagram

The artifact which should solve the identified problem was developed using CodeIgniter and Python Flask. CodeIgniter (CI) framework was used to create a web-based application that was compiled using the PHP language [45]. Over the years, PHP has developed from a basic web scripting language to becoming one of the most extensively used languages for building large web-based applications and frameworks, both in the open-source community and in industry [46]. For the development of data mining reports in this research, the Python Flask web application framework was used, which provides various services such as built-in HTTP server, support for unit testing, and RESTful web service.

3.1.4 Demonstration

This step involves demonstration and initial evaluation of the design and developed artifact. This is a three phases cycle comprising of review of objectives, design and demonstration to assess the relevance of the proposed solution to the environment.

3.1.5 Evaluation

This is the step where the evaluation is done after building the prototype or the artifact, it needs to be evaluated against predefined evaluation criteria. Here the artifact is appraised to determine its worthiness in how it solve the problem identified. Therefore, all the tests conducted and strategies used are now explained. Evaluation also involves checking whether what was expected of the solution and its objectives were done.

3.1.6 Communication.

This is the final phase in DSR and involves expert review and refining the artifact. It also involves contribution to theory and methodology and practice. This contribution can be in form of paper publication.

3.2 Data Mining Model Development

Data Mining is a technique used to extract useful information or patterns from unprocessed data. It is utilized in various applications such as political model forecasting, weather pattern model forecasting, website ranking forecasting, and more [47]. The development of a data mining model typically consists of five primary steps, as illustrated in Figure 5.

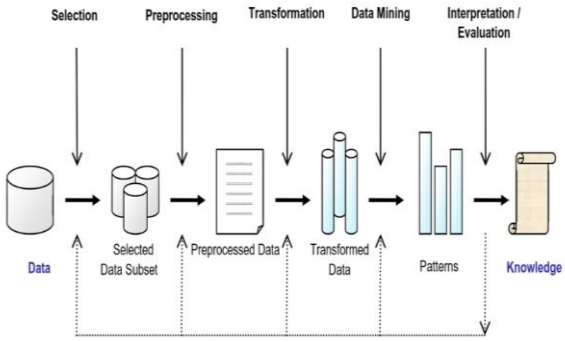

Figure 5 illustrates the five main steps involved in the Data Mining process [48].

3.2.1 Data selection

Datasets are crucial for the advancement of machine learning research, as they serve as the basis for the models we create and implement, as well as our primary means of benchmarking and evaluating their performance [49]. The dataset for diseases was generated based on WHO and Ministry of Health reports while patient's names were randomly generated.

3.2.2 Data pre-processing

Once the focus group data was obtained, it underwent a cleaning process to ensure that the information was accurate and formatted correctly [24].

3.2.3 Data Transformation

The next step in the data analysis was the ETL process, which involved extracting data from the dataset and transforming it into a format suitable for the mining procedure [24]. This process included selecting relevant data for feature selection, which was done using the following Python code to prepare and transform the loaded data.

```
def prepare_inputs(X_train, X_test):
    oe = OrdinalEncoder()
    oe.fit(X_train)
    X_train_enc = oe.transform(X_train)
    X_test_enc = oe.transform(X_test)
    return X_train_enc, X_test_enc
```

3.2.4 Data mining

Data mining techniques were employed in this study to extract valuable insights from the raw data and identify significant patterns. Python was used in this study and Table 1 below presents the algorithms that were used in this study.

Table 1. The List of Algorithms used

Techniques	**Algorithm**
Feature selection	• LabelEncoder from sklearn.preprocessing • OrdinalEncoder from sklearn.preprocessing • SelectKBest from sklearn.feature_selection • chi2 from sklearn.feature_selection
Data visualization	• from seaborn as sns • from matplotlib.pyplot as plt • from matplotlib.figure import Figure • from matplotlib.backends.backend_agg import FigureCanvasAgg as FigureCanvas
Association	• fpgrowth from mlxtend.frequent_patterns • apriori from mlxtend.frequent_patterns
Classification	• "from sklearn.model_selection import train_test_split • from sklearn.model_selection import cross_val_score • from sklearn.model_selection import StratifiedKFold • from sklearn.linear_model import LogisticRegression • from sklearn.tree import DecisionTreeClassifier • from sklearn.neighbors import KNeighborsClassifier • from sklearn.discriminant_analysis import LinearDiscriminantAnalysis • from sklearn.naive_bayes import GaussianNB • from sklearn.model_selection import train_test_split • from sklearn.metrics import classification_report • from sklearn.metrics import confusion_matrix • from sklearn.metrics import accuracy_score • from sklearn.svm import SVC"

4.2.5 Interpretation and evaluation

The results were analysed and explicit knowledge was created by visualizing the data through reports and dashboards [24]. The Interpretation and evaluation are presented in the next section.

4. Results

4.1 *Web-based ehealth*

4.1.1 *CRUD (*Create, Read, Update, and Delete)

CRUD methods were developed for the web application for the following are the tables used in the database:

- Staff
- Patients
- Outpatient
- Inpatient
- birth record
- Hivtesting
- Covidtesting
- X-ray
- medicines
- death record
- ward
- observation
- antenatal
- adultpeadiatricart
- Family planning
- Under five
- treatment

Fig 6 below shows the CRUD method for the patient's table.

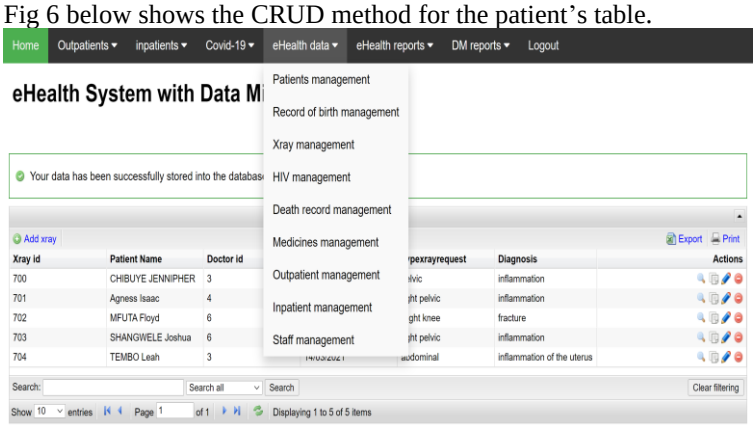

Fig 6 CRUD for patients table

4.1.2 Web Technology

CodeIgniter (CI) is a framework was used to create a web-based application that is compiled using the PHP language. In CI there are numerous classes that form Library and helper.

4.1.3 Reports

The system is able to produce a number of e-health reports (see Fig 7).

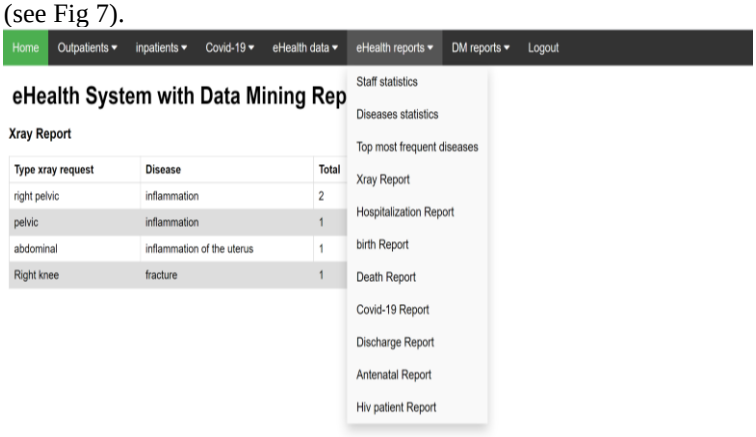

Fig 7. ehealth reports

The following outlines the types of reports being generated from the system.

- The Xray report shows the type of x-ray request, disease and totals.
- Hospitalization report contains name of the ward where the patient is admitted, diagnosis, the date and time of admission and total number of patients. It also includes the type of disease that the patients are diagnosed with in each ward.
- Birth report shows date, birth weight and totals including the number of babies born with the same weight.
- Death Report which indicates the number of deaths, disease, and associated health facility ward.
- Covid 19 Report indicates the Covid 19 status and associated symptoms.
- The Discharge report specifies the date of discharge, the disease, and the specific ward.
- The antenatal Report specifies the health facility name and the number of expectant mothers.
- HIV Patient Report indicates the patient status results and whether the patient is on treatment or not.

4.2 Data mining report

4.2.1 Descriptive statistics

The descriptive statistics of the data for a one year is provided in the table 2 below

Table 2 Descriptive statistics

	systolic	diastolic	pulse	temperature	weight	age
count	22794.000000	22794.000000	22794.000	22794.000	22794.00	22794.000000
mean	139.805870	85.063657	70.087567	37.049838	64.639598	50.274716
std	35.050564	20.601273	17.598494	1.454850	14.376817	4.941553
min	80.000000	50.000000	40.000000	35.000000	40.000000	41.831600
25%	109.000000	67.000000	55.000000	36.000000	52.000000	45.980175
50%	140.000000	85.000000	70.000000	37.000000	65.000000	50.228600
75%	170.000000	103.000000	85.000000	38.000000	77.000000	54.581800
max	200.000000	120.000000	100.000000	40.000000	90.000000	58.830900

4.2.2 Prevalence and incidence diseases

Table 3 below indicates the top 10 prevalence and incidence diseases from the data.

Table 3. Top 10 Prevalence and incidence diseases

	Diagnosis	prevalence	incidence
0	HIV/AIDS	0.0251	0.1403
1	Neonatal	0.0192	0.1073
2	Hypertension	0.0123	0.0688
3	Tuberculosis	0.0103	0.0578
4	Ischemic heart disease	0.0094	0.0523
5	Diarrheal diseases	0.0089	0.0495
6	Cirrhosis	0.0069	0.0385
7	Malaria	0.0069	0.0385
8	Vitiligo	0.0064	0.0358
9	Severe Migraine	0.0064	0.0358
10	T.B Meningitis	0.0059	0.0330

4.2.3 Feature selection

The results of feature selection are presented in the table 4 below. It was observed that pulse rate and age were the most significant the resultset.

Table 4 Feature selection

Variable	Feature	Value
Pulse	Feature 0	106.037552
temperature	Feature 1	4.390025
Weight	Feature 2	0.685386
Age	Feature 3	68.052073
place of birth	Feature 4	3.249173
Compound	Feature 5	0.142780
District	Feature 6	0.380944
Gender	Feature 7	1.185059

4.2.4 Data visualization

The data mining component of the web application was able to translate the large data sets and metrics into charts, graphs and other visuals. Data visualization was able to makes it easier for stakeholders to identify and share trends, outliers, and new insights about the information represented by the data. Fig 8 and Fig 9 provides two examples of graphs that were produced by the web application.

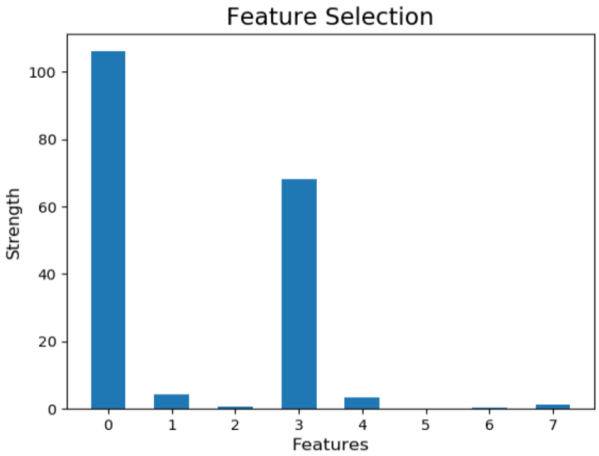

Fig 8 Feature Selection

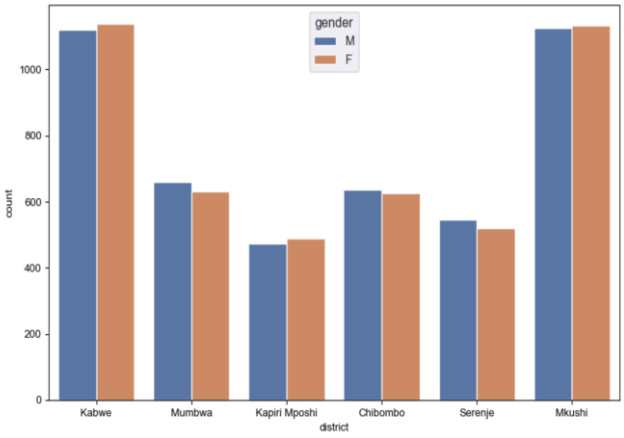

Fig 9 number of patients by district

4.2.5 Association

The results of FP growth for minimum support 60% is given in the table 5 showing high association for HIV/AIDS followed by Neonatal.

Table 5 FP growth

	support	itemsets

0	0.918860	(HIV/AIDS)
1	0.809211	(Neonatal)
2	0.666667	(Hypertension)
3	0.750000	(HIV/AIDS, Neonatal)
4	0.660088	(Hypertension, Neonatal)
5	0.618421	(HIV/AIDS, Hypertension)
6	0.611842	(HIV/AIDS, Hypertension, Neonatal)

The results of the Apriori for minimum support 60% is given in the table 6.

Table 6 Appriori

	support	**itemsets**
0	0.918860	(HIV/AIDS)
1	0.666667	(Hypertension)
2	0.809211	(Neonatal)
3	0.618421	(HIV/AIDS, Hypertension)
4	0.750000	(HIV/AIDS, Neonatal)
5	0.660088	(Hypertension, Neonatal)
6	0.611842	(HIV/AIDS, Hypertension, Neonatal)

The table 7 provides the Appriori Rules and the values.

4.2.6 Classification

The Correlation Matrix of the dataset is provided in Table 8.

Table 8 Correlation Matrix

	systolic	**diastolic**	**pulse**	**temperature**	**weight**	**age**
systolic	1.000000	-0.007701	0.001937	0.008978	0.014006	0.004397
diastolic	-0.007701	1.000000	-0.011497	-0.004067	-0.009367	-0.007251
pulse	0.001937	-0.011497	1.000000	-0.008745	-0.001525	-0.004472
temperature	0.008978	-0.004067	-0.008745	1.000000	0.004909	0.004094
weight	0.014006	-0.009367	-0.001525	0.004909	1.000000	-0.006555
age	0.004397	-0.007251	-0.004472	0.004094	-0.006555	1.000000

Table 7 Appriori Rules

	antecedents	**consequents**	**antecedent support**	**consequent support**	**support**	**confidence**	**lift**	**leverage**	**conviction**
4	(Hypertension)	(Neonatal)	0.666667	0.809211	0.660088	0.990132	1.223577	0.120614	19.333333
6	(HIV/AIDS, Hypertension)	(Neonatal)	0.618421	0.809211	0.611842	0.989362	1.222626	0.111409	17.934211
1	(Hypertension)	(HIV/AIDS)	0.666667	0.918860	0.618421	0.927632	1.009547	0.005848	1.121212
8	(Hypertension, Neonatal)	(HIV/AIDS)	0.660088	0.918860	0.611842	0.926910	1.008762	0.005314	1.110148
3	(Neonatal)	(HIV/AIDS)	0.809211	0.918860	0.750000	0.926829	1.008673	0.006449	1.108918
10	(Hypertension)	(HIV/AIDS, Neonatal)	0.666667	0.750000	0.611842	0.917763	1.223684	0.111842	3.040000
2	(HIV/AIDS)	(Neonatal)	0.918860	0.809211	0.750000	0.816229	1.008673	0.006449	1.038192
7	(HIV Neonatal)								
5	(Neonatal)	(Hypertension)	0.809211	0.666667	0.660088	0.815718	1.223577	0.120614	1.808824
11	(Neonatal)	(HIV/AIDS, Hypertension)	0.809211	0.618421	0.611842	0.756098	1.222626	0.111409	1.564474
0	(HIV/AIDS)	(Hypertension)	0.918860	0.666667	0.618421	0.673031	1.009547	0.005848	1.019465
9	(HIV/AIDS)	(Hypertension, Neonatal)	0.918860	0.660088	0.611842	0.665871	1.008762	0.005314	1.017309

Table 9 provides the Classification report

	precision	**recall**	**f1-score**	**support**
Cirrhosis	0.000000	0.000000	0.000000	243.000000
Diarrheal diseases	0.000000	0.000000	0.000000	443.000000
HIV/AIDS	0.242098	1.000000	0.389822	1103.000000
Hypertension	1.000000	0.002262	0.004515	442.000000
Ischemic heart disease	0.000000	0.000000	0.000000	334.000000
Lower respiratory infection	0.000000	0.000000	0.000000	242.000000
Neonatal	1.000000	0.002710	0.005405	738.000000
Severe Migraine	0.000000	0.000000	0.000000	218.000000
T.B Meningitis	0.000000	0.000000	0.000000	203.000000
Tuberculosis	0.000000	0.000000	0.000000	360.000000
Vitiligo	0.000000	0.000000	0.000000	233.000000
accuracy	0.242597	0.242597	0.242597	0.242597
macro avg	0.203827	0.091361	0.036340	4559.000000
weighted avg	0.317402	0.242597	0.095626	4559.000000

Table 10 Confusion matrix

	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
0	0	0	243	0	0	0	0	0	0	0	0
1	0	0	443	0	0	0	0	0	0	0	0
2	0	0	1103	0	0	0	0	0	0	0	0
3	0	0	441	1	0	0	0	0	0	0	0
4	0	0	334	0	0	0	0	0	0	0	0
5	0	0	242	0	0	0	0	0	0	0	0
6	0	0	736	0	0	0	2	0	0	0	0
7	0	0	218	0	0	0	0	0	0	0	0
8	0	0	203	0	0	0	0	0	0	0	0
9	0	0	360	0	0	0	0	0	0	0	0
10	0	0	233	0	0	0	0	0	0	0	0

5. Discussion

5.1 Web-based e-health

The first objective of this research was to develop a web-based e-health system for Zambian health centres. This was achieved through the design of a web-based system which can be accessed online. The central location of reports and statistics provides continuity in flow of information in health facilities for informed decision making.

Compared to SmartCare and Zambia Health Management Information System, the designed ehealth system is unique in terms of its approach to accessing and storing information. Since it is a web-based system, information can be accessed from anywhere at any time, and it is centralized. SmartCare Health Information System uses client care cards where individual's health information is stored as a way of maintaining continuousness of care between visits, health services and health facilities.

Each patient is equipped with a SmartCare card embedded with a chip that contains all their health information, as well as a copy of the information stored at the clinic. In a web-

based system, patient information is readily available and can be accessed by any health facility through a browser unlike using a SmartCare card that is limited to only Health facilities that have SmartCare system and with card readers that are functional.

The DHIS2 use pivot tables and values in the pivot tables are non-editable and all the names and numbers are fetched directly from the DHIS2 database and to edit the contents of a pivot table the content is copied to a normal spreadsheet. "A table with around 1 million values (rows of data) tend to become less responsive to updates (refresh) and pivoting operations". The proposed solution eliminates the need for pivot tables and the DHIS2 database since all data is stored in a centralized database and data mining tools can directly retrieve the data from the database.

5.2 Important Data mining tools and techniques

The second objective of this study was to evaluate critical data mining tools and techniques for e-health in Zambia. This was achieved through literature review and implementation of some of the tools and techniques which were relevant to the study. Among the notable techniques include Feature selection, Data visualization, Association and Classification.

Table 11 shows some of the important data mining tools and techniques identified for e-health in Zambia during the study.

Table 11 Data mining tools

Algorithm	**DM tools and techniques**	**Ehealth application**	**References**
Decision tree	Classification and Clustering	detection of diabetes, digestion, and kidney diseases	[37]
Random Forest Classifier	Classification	death risk prediction	[39]
"Decision Tree" Classifier, "Gaussian Naive Bayes" Classifier, "Random Forest" Classifier," Logistic Regression" Classifier, "k-Nearest" Neighbour Classifier," Multilayer Perceptron" Classifier, "Support Vector Machine" Classifier	Classification	Predicting Dengue, Diabetes and Thyroid	[40]

5.3 Integration of Data Mining in web-based e-health

The third objective of this study was to integrate important data mining tools and techniques in the web-based e-health system for Zambia. This objective was achieved by using techniques such as Feature selection, Data visualization, Association, Classification to analyse and extract important information from the dataset to generate data mining reports and statistics.

Data mining techniques are a powerful tool when incorporated with web-based systems especially when working with big data. The generated reports help to meet user's requirements as well aiding decision making. The e-health web system was validated by use of questionnaires that were distributed to various health facilities. The information that was obtained through validation was analysed to determine system effectiveness and efficiency.

6. Conclusion

In this research, we have presented the problem of SmartCare in terms of SmartCare cards. The application framework that has been is able to solve the identified SmartCare problems such as making the information readily available all the time and easily accessible. The framework is user-friendly, provides precise information, and is exceptional in terms of data uploading and patient data analysis.

This research contributed an artifact in form of web based e-health system with data mining reporting. The e-health system has the potential of improving efficiency and service delivery in the health sector. It will be easy for the management in health facilities to make informed decisions since there will be continuous flow of information. There will be no mismanagement of patients since the history of the patient will be followed. Additionally, management of some outbreaks in certain compounds will be easy to manage since the system will be able to show the locations where patients are coming from. Furthermore, data mining reports will be generated to show prevalence and incidence diseases and the Ministry of Health can use these reports for budgeting purposes. One limitation of this study is that the validation of the data mining reports relied on a dataset that included some randomly generated data.

This research used feature selection, data visualization, association and classification as part of the data mining reporting. Future research can include clustering, machine learning, neural networks and outlier detection. Information on patients accessing renal unit services should be included to know how many patients are on dialysis treatment. This will help the health sector to plan and allocate enough funds since the treatment is very expensive.

References

[1] C. Butpheng, K.-H. Yeh, and H. Xiong, 'Security and Privacy in IoT-Cloud-Based e-Health Systems—A Comprehensive Review', *Symmetry*, vol. 12, no. 7, p. 1191, Jul. 2020, doi: 10.3390/sym12071191.

[2] P. C. Cheruiyot, 'Influence of selected strategies on implementation of donor assisted e-health management systems in Kenya: a survey of public health facilities in Nakuru County', Kabarak University, 2018.

[3] T. Chawurura, R. Manhibi, J. van Dijk, and G. van Stam, 'eHealth in Zimbabwe, a case of techno-social development', in *International Conference on Social Implications of Computers in Developing Countries*, 2019, pp. 15–26.

[4] D. Leza and J. Phiri, 'Challenges of medical records interoperability in developing countries: a case study of the University teaching hospital in Zambia', *Int. J. Adv. Comput. Sci. Appl.*, vol. 10, no. 1, 2019.

[5] Ministry of Health, Zambia, 'e-health strategy 2017-2021'. Ministry of Health, Apr. 28, 2017. [Online]. Available: https://www.moh.gov.zm/?wpfb_dl=89

[6] W. M. Haddad, J. M. Bailey, B. Gholami, and A. R. Tannenbaum, 'Clinical decision support and closed-loop control for intensive care unit sedation', *Asian J. Control*, vol. 15, no. 2, pp. 317–339, 2013.

[7] M. van Reisen, 'International cooperation in the digital era', Leiden, 2017.

[8] K. AP, B. UJ, and N. EB, 'Health information systems in developing countries: Case of African countries', 2020.

[9] M. A. Zijlman, 'The opportunities for patient identification and E-Health in Africa: a study into the patient identification problems and opportunities for E-Health for the Lamin Health Center in The Gambia', University of Twente, 2020.

[10] K. Svalestuen, 'Reappropriation of Generic Information Systems', 2018.

[11] J. Phiri and M. Nyirenda, 'Medical Equipment and Laboratory Services Support System', *researchgate.net*, 2019.

[12] B. Yang, Z. Zhao, and J. Ma, 'Marine accidents analysis based on data mining using K-medoids clustering and improved A priori algorithm', in *IOP Conference Series: Earth and Environmental Science*, 2018, vol. 189, no. 4, p. 042006.

[13] R. C. Chellah and D. Kunda, 'An assessment of factors that affect the implementation of big data analytics in the Zambian health sector for strategic planning and predictive analysis: a case of Copperbelt province', *Int. J. Electron. Healthc.*, vol. 11, no. 2, p. 101, 2020, doi: 10.1504/IJEH.2020.113196.

[14] T. Mageto and D. Neagu, 'Design and Development of E-Health System', *J. Comput. Sci. Appl.*, vol. 6, no. 1, pp. 1–16, 2018.

[15] K. M. Haakalaki, J. Phiri, and M. K. Kabemba, 'A Model for an Electronic Health Information Management System with Structural Interoperability in Heterogeneous Environments for continued Health Care', *Zamb. ICT J.*, vol. 2, no. 2, pp. 28–35, 2018.

[16] G. Biemba *et al.*, 'A mobile-based community health management information system for community health workers and their supervisors in 2 districts of Zambia', *Glob. Health Sci. Pract.*, vol. 5, no. 3, pp. 486–494, 2017.

[17] M. Gregory and S. Tembo, 'Implementation of E-health in developing countries challenges and opportunities: a case of Zambia', *Sci. Technol.*, vol. 7, no. 2, pp. 41–53, 2017.

[18] B. Mutabazi, 'A case study to investigate the challenges of EMR implementation in four district Hospitals in Rwanda', University of Rwanda, 2016.

[19] M. R. Hoque, Y. Bao, and G. Sorwar, 'Investigating factors influencing the adoption of e-Health in developing countries: A patient's perspective', *Inform. Health Soc. Care*, vol. 42, no. 1, pp. 1–17, 2017.

[20] J. Braa and S. Sahay, 'The DHIS2 open source software platform: evolution over time and space', *LF Celi Glob. Health Inform.*, vol. 451, 2017.

[21] E. Adu-Gyamfi, P. Nielsen, and J. I. Sæbø, 'The dynamics of a global health information systems research and implementation project', in *SHI 2019. Proceedings of the 17th Scandinavian Conference on Health Informatics, November 12-13, 2019, Oslo, Norway*, 2019, no. 161, pp. 73–79.

[22] A. Farnham, J. Utzinger, A. V. Kulinkina, and M. S. Winkler, 'Using district health information to monitor sustainable development', *Bull. World Health Organ.*, vol. 98, no. 1, p. 69, 2020.

[23] P. Saunders-Hastings, 'DHIS2 as a tool for health impact assessment in low-resource settings: examining opportunities for expanding use of routine health data', in *Proceedings of the 38th annual conference of the international association for impact assessment. Durban, South Africa*, 2018, pp. 16–19.

[24] R. Ray, 'Advances in data mining: Healthcare applications', *Int. Res. J. Eng. Technol. IRJET*, vol. 5, no. 03, pp. 2395–0056, 2018.

[25] M. Saqlain, W. Hussain, N. A. Saqib, and M. A. Khan, 'Identification of heart failure by using unstructured data of cardiac patients', in *2016 45th International Conference on Parallel Processing Workshops (ICPPW)*, 2016, pp. 426–431.

[26] N. Chikshe, T. Dixit, R. Gore, and P. Akade, 'Hybrid approach for heart disease detection using clustering and ANN', *Int. J. Recent Innov. Trends Comput. Commun.*, vol. 4, no. 1, pp. 119–122, 2016.

[27] M. R. Sengamuthu, M. R. Abirami, and D. Karthik, 'Various data mining techniques analysis to predict diabetes mellitus', *Int Res J Eng TechnolIRJET*, vol. 5, no. 05, 2018.

[28] J. Santos-Pereira, L. Gruenwald, and J. Bernardino, 'Top data mining tools for the healthcare industry', *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 34, no. 8, pp. 4968–4982, Sep. 2022, doi: 10.1016/j.jksuci.2021.06.002.

[29] D. Sharma, A. Sharma, and V. Mansotra, 'A Literature Survey on Data Mining Techniques to Predict Lifestyle Diseases', *Int. J. Res. Appl. Sci. Eng. Technol. IJRASET*, vol. 5, pp. 1575–1576, 2017.

[30] V. Rogeith and S. Magesh, 'A Survey On Health Care Data Using Data Mining Techniques', *Int. J. Pure Appl. Math.*, vol. 117, no. 16, pp. 665–672, 2017.

[31] L. C. Borges, V. M. Marques, and J. Bernardino, 'Comparison of data mining techniques and tools for data classification', in *Proceedings of the International C* Conference on Computer Science and Software Engineering*, 2013, pp. 113–116.

[32] A. M. Khedr, Z. Al Aghbari, A. Al Ali, and M. Eljamil, 'An efficient association rule mining from distributed medical databases for predicting heart diseases', *IEEE Access*, vol. 9, pp. 15320–15333, 2021.

[33] X. Gao, F. Q. Xu, and Z. M. Zhu, 'The application of improved FP-growth algorithm in disease complications', in *Advanced Science and Industry Research Center. Proceedings of 2019 International Conference on Computational Modeling, Simulation and Optimization (CMSO 2019)*, 2019, pp. 118–122.

[34] D. S. Ginting, H. Mawengkang, and S. Efendi, 'Modification of A priori Algorithm focused on confidence value to association rules', in *IOP Conference Series: Materials Science and Engineering*, 2018, vol. 420, no. 1, p. 012125.

[35] T. Andi and E. Utami, 'Association rule algorithm with FP growth for book search', in *IOP Conference Series: Materials Science and Engineering*, 2018, vol. 434, no. 1, p. 012035.
[36] M. M. M. Pizzichini, C. M. Patino, and J. C. Ferreira, 'Measures of frequency: calculating prevalence and incidence in the era of COVID-19', *J. Bras. Pneumol.*, vol. 46, 2020.
[37] S. Nagarajan and R. M. Chandrasekaran, 'Design and implementation of expert clinical system for diagnosing diabetes using data mining techniques', *Indian J. Sci. Technol.*, vol. 8, no. 8, pp. 771–6, 2015.
[38] A. MAHAR and S. RAJPAR, 'Data Mining Techniques in E-Health Systems: An Analysis', *Sindh Univ. Res. J.-SURJ Sci. Ser.*, vol. 49, no. 2, 2017.
[39] R. Valter, S. Santiago, R. Ramos, M. Oliveira, L. O. M. Andrade, and I. C. de HC Barreto, 'Data mining and risk analysis supporting decision in brazilian public health systems', in *2019 IEEE International Conference on E-health Networking, Application & Services (HealthCom)*, 2019, pp. 1–6.
[40] Md. S. R. Zishan, M. A. Mohamed, C. A. Hossain, R. Ahasan, and S. M. Sharun, 'Design and Deployment of E-Health System Using Machine Learning in the Perspective of Developing Countries':, *Int. J. Ambient Comput. Intell.*, vol. 13, no. 1, pp. 1–20, Jan. 2022, doi: 10.4018/IJACI.293186.
[41] A. Hevner and S. Chatterjee, 'Design science research in information systems', in *Design research in information systems*, Springer, 2010, pp. 9–22.
[42] R. Baskerville, A. Baiyere, S. Gregor, A. Hevner, and M. Rossi, 'Design science research contributions: Finding a balance between artifact and theory', *J. Assoc. Inf. Syst.*, vol. 19, no. 5, p. 3, 2018.
[43] R. Ismail *et al.*, 'Employing Method for a Method: Design and Evaluation of the Proposed PDEduGame Process Method using Design Science Research Method', in *Proceedings of the 2020 6th International Conference on Computer and Technology Applications*, 2020, pp. 87–91.
[44] W. H. Organization, *WHO guideline: recommendations on digital interventions for health system strengthening*. World Health Organization, 2019.
[45] R. N. Thakur and U. S. Pandey, 'The role of model-view controller in object oriented software development', *Nepal J. Multidiscip. Res.*, vol. 2, no. 2, pp. 1–6, 2019.
[46] A. Siame and D. Kunda, 'Evolution of PHP Applications: A Systematic Literature Review', *Int. J. Recent Contrib. Eng. Sci. IT IJES*, vol. 5, no. 1, p. 28, Mar. 2017, doi: 10.3991/ijes.v5i1.6437.
[47] Priya Pedamkar, 'Models in Data Mining | Techniques | Algorithms | Types', *EDUCBA*, Dec. 18, 2019. https://www.educba.com/models-in-data-mining/ (accessed Nov. 21, 2022).
[48] I. Kavakiotis, O. Tsave, A. Salifoglou, N. Maglaveras, I. Vlahavas, and I. Chouvarda, 'Machine learning and data mining methods in diabetes research', *Comput. Struct. Biotechnol. J.*, vol. 15, pp. 104–116, 2017.
[49] A. Paullada, I. D. Raji, E. M. Bender, E. Denton, and A. Hanna, 'Data and its (dis) contents: A survey of dataset development and use in machine learning research', *Patterns*, vol. 2, no. 11, p. 100336, 2021.