ГОДИШНИК
на Софийския университет „Св. Климент Охридски“

Факултет по математика и информатика

ANNUAL
of Sofia University “St. Kliment Ohridski”

Faculty of Mathematics and Informatics

Том/Volume 110

УНИВЕРСИТЕТСКО ИЗДАТЕЛСТВО „СВ. КЛИМЕНТ ОХРИДСКИ“
ST. KLIMENT OHRIDSKI UNIVERSITY PRESS
СОФИЯ • 2023 • SOFIA

Годишник на Софийския университет „Св. Климент Охридски“
Факултет по математика и информатика

Annual of Sofia University "St. Kliment Ohridski"
Faculty of Mathematics and Informatics

Address for correspondence

Faculty of Mathematics and Informatics
Sofia University "St. Kliment Ohridski"
5 J. Bourchier Blvd., POB 48
1164 Sofia, BULGARIA
email: annuaire@fmi.uni-sofia.bg

Aims and Scope. The *Annual* is the oldest Bulgarian journal, founded in 1904, devoted to pure and applied mathematics, mechanics and computer science. It is reviewed by *Zentralblatt für Mathematik*, *Mathematical Reviews* and the Russian *Referativnii Jurnal*. The *Annual* publishes significant and original research papers of authors both from Bulgaria and abroad in some selected areas that comply with the traditional scientific interests of the Faculty of Mathematics and Informatics at the Sofia University "St. Kliment Ohridski", i.e., algebra, geometry and topology, analysis, probability and statistics, mathematical logic, theory of approximations, numerical methods, computer science, classical, fluid and solid mechanics, and their fundamental applications.

ISSN 1313–9215 (Print)
ISSN 2603–5529 (Online)

CONTENTS

Editorial 5

Simeon Abanos, Evgeniy Krastev, Petko Kovachev and Dimitar Tcharaktchiev. Modeling and clinical data exchange in registration of infectious disease cases 7

Kristina Arnaoudova and Maria Nisheva-Pavlova. Enterprise data and semantic modelling: conceptual model of information technology incident management 25

Melania Berbatova and Filip Ivanov. An improved Bulgarian natural language processing pipeline 37

Natasha Blazeska-Tabakovska, Snezana Savoska, Andrijana Bocevska, Ilija Jolevski, Blagoj Ristevski and Vladimir Trajkovik. Increase digital health literacy with the practical implementation of an e-learning system according to WCAG compliance standards 51

Vladimir Dimitrov. CAPEC ontology 63

Mohamad Fakhredine and Teodora Bakardjieva. Machine learning for biomedical-focus on brain signals 85

Oksana Kharchenko and Zlatinka Kovacheva. Frequency modulated signal standing out by stochastic resonance effect 95

Martin Kostov and Kalinka Kaloyanova. Real-time data integration in information systems using stream processing for medical data 101

Mariya Miteva and Maria Nisheva-Pavlova. Preprocessing techniques for brain MRI scans: a comparative analysis for radiogenomics applications 111

Mitko Trpkoski, Andrijana Bocevska, Igor Nedelkovski and Snezana Savoska. Game theme based instructional module to teach loops and choice statements in computer science courses 127

Elitsa Venchova and Vasil Georgiev. Next generation service models of mobile autonomous assistants 139

Editorial

The *International Conference on Information Systems, Embedded Systems, and Intelligent Applications* (*ISESIA*) 2023 was held in Sofia, Bulgaria, from May 26 to May 27, 2023. You can access detailed information about the event at ISESIA's official website `https://isesia.fmi.uni-sofia.bg/`.

This conference was a collaborative undertaking between the Department of Computer Informatics at the Faculty of Mathematics and Informatics of Sofia University "St. Kliment Ohridski" and the Bulgarian Chapter of the Association for Information Systems (BulAIS), with financial support from both the Science Fund of Sofia University and BulAIS.

Building on the legacy of 15 successful editions of the Information Systems and Grid Technologies (ISGT) conference, ISESIA 2023 retained its core focus on addressing fundamental challenges and technological advancements in the domain of Information Systems (IS) while expanding its thematic horizons. The conference's primary objective was to foster an open and dynamic platform for the exchange of ideas and the exploration of ongoing research trends, innovative methodologies, and original research outcomes across various facets of Information Systems, Intelligent Systems, and Distributed and Cloud Systems.

ISESIA 2023 encompassed a wide spectrum of interconnected topics, including:

- Databases and Information Systems Integration
- Big Data Innovations
- Data Analytics
- IS Security, Privacy and Ethics
- Business Models and Business Processes
- Advances in IS Programs and Education
- Knowledge Representation and Reasoning
- Machine Learning
- Data Mining and Knowledge Discovery
- Natural Language Processing and Information Retrieval
- Responsible Artificial Intelligence
- Smart Robots
- Computer Vision

- Modeling, Clustering, Virtualization, SOA
- P2P, IoT, Mobility
- Application-specific Grid and Cloud.

In conjunction with the central ISESIA program, the conference hosted three specialized workshops:

- *MIRACle*: Modeling Robotic Assistance
- *eHealth*: IS and Health Information Technology Adoption
- *UNITe*: Data Engineering, Data Analysis and Visualization

The *MIRACle workshop* concentrated on the latest advancements in the fields of embedded systems, mobile robotics, and autonomous human assistance. Key research areas covered in this workshop included: advancements in high-level and intelligent control techniques for mobile robotics; scalable multi-agent, peer-to-peer, and distributed embedded systems; innovative data services for sensor streaming and related topics.

The primary objective of the *eHealth workshop* was to address key issues and offer a comprehensive perspective on the contemporary implementation and adoption of eHealth. This workshop featured works related to various aspects of eHealth, such as artificial intelligence in healthcare, digital health literacy, and the application of information technologies in the healthcare sector.

The *UNITe workshop* aimed to provide a platform for the exchange of ideas and results in thematic areas related to data engineering, effective data utilization, and the creation of efficient models for data analysis and visualization. Research presented at this workshop primarily focused on new approaches for semantic modeling, the development of decision support systems, and enhancing the utilization of available data by fostering interoperability at different levels.

A total of 37 papers were submitted for consideration at ISESIA 2023. Each paper underwent review by at least two members of the Program Committee and a panel of invited experts. Ultimately, the Program Committee selected 29 papers for presentation at the conference, with 11 of them accepted for publication in the current volume of the Annual of Sofia University, Faculty of Mathematics and Informatics.

ISESIA 2023 featured a diverse range of research and practical results, shared experiences, and in-depth discussions, offering significant value to the Information Systems community, particularly benefiting young IS researchers and PhD students.

ГОДИШНИК НА СОФИЙСКИЯ УНИВЕРСИТЕТ „СВ. КЛИМЕНТ ОХРИДСКИ“
ФАКУЛТЕТ ПО МАТЕМАТИКА И ИНФОРМАТИКА
Том 110

ANNUAL OF SOFIA UNIVERSITY “ST. KLIMENT OHRIDSKI”
FACULTY OF MATHEMATICS AND INFORMATICS
Volume 110

MODELING AND CLINICAL DATA EXCHANGE IN REGISTRATION OF INFECTIOUS DISEASE CASES

SIMEON ABANOS, EVGENIY KRASTEV,
PETKO KOVACHEV AND DIMITAR TCHARAKTCHIEV

The registration of communicable disease cases aims to limit their prevalence. The Ordinance on the Procedure for Registration, Notification and Reporting on Communicable Diseases, issued by the Ministry of Health, determines the rules for the implementation of the registration procedure in Bulgaria. Currently, this information is communicated by phone, email or on paper, and then the data is manually re-entered. This significantly complicates the exchange of information between health care participants. In addition, valuable information to produce reports noting disease prevalence levels is lost or fragmented. This paper aims to resolve these problems by the means of information technologies and ensure efficient storage, processing and management of data registration for patients with communicable diseases. The business process for the implementation has been studied in detail. The main roles, resources and sequence of activities at each stage of the application of the regulation have been determined. A relational database model is created for centralized storage, processing and management of data for registered patients and their contacts. An algorithm is developed and implemented for the automated determination of cases based on the symptoms of each disease specified in the regulation. ICD-10 is used, thus, creating the possibility of exchanging the collected data with international centers for tracking the spread of the disease. A three-tier application is created where web services perform the activities of registration, communication and reporting of communicable cases. The obtained results were tested with real data, consulted with experts from the health inspections and demonstrate the advantages of the developed software for fulfilling the requirements of the provisions in national legislation.

Keywords: eHealth, interoperability, clinical data exchange, clinical information models, communicable diseases, disease prevention

CCS Concepts:

• Software and its engineering~Software creation and management~Designing software~Software design engineering;

• Information systems~Data management systems~Information integration~Data exchange

DOI: 10.60063/GSU.FMI.110.7-23

1. Introduction

Electronic health (eHealth) is defined by the World Health Organization (WHO) as "cost-effective and secure use of information and communications technologies in support of health and health-related fields, including health-care services, health surveillance, health literature, and health education, knowledge and research" [11]. This includes the exchange of a wide range of health information, including the use of electronic health records (EHR) [5], telemedicine [3], digital tools and patient information management and exchange systems. eHealth has the potential to revolutionize the way health care is delivered, offering patients a more efficient, accessible and personalized system. The integration of technology, particularly the Internet and electronic devices, into the health care system enables instant communication and the exchange of large volumes of information between health care providers and patients. Using the capabilities of these technologies, eHealth provides patients with access to health services remotely and at any time through the use electronic devises; enables the automation of many healthcare processes, reducing the time and resources needed to provide care; gives patients greater access to their health information, allowing them to take a more active role in their health care and make informed decisions about their health; allows healthcare provides access to a more complete and accurate picture of a patient's health, leading to more informed decision-making and better health outcomes; has the potential to optimize healthcare costs by reducing the need for expensive and time-consuming procedures, such as duplicate tests and unnecessary hospital visits.

The lack of standardization in eHealth systems can lead to difficulties in the exchange of information, resulting in fragmented patient records, duplicated tests and procedures, increasing the work of healthcare specialists, respectively reducing the time they could devote to the patient and an overall deterioration of the quality of health care. The use of heterogeneous and specific software solutions and systems by different healthcare organizations can also pose challenges in data exchange, leading to difficulties, fragmentation and a lack of access to important information.

Despite its many potential benefits, the implementation of eHealth is not without its challenges. One of the most significant of which is the lack of interoperability [4]. It is the ability of different electronic systems and devices to communicate with each other and to exchange and receive information, understanding it in its semantic context. In the healthcare industry, interoperability is critical to delivering high-quality patient care and improving the healthcare experience. The goal of interoperability is to enable native exchange of health information between different systems and among healthcare participants, such as hospitals, clinics, health insurance companies and patients, without loss of data meaning and functionality. In the context of eHealth, interoperability can be divided into several levels, among them, the most important for efficient data exchange is semantic interoperability [6]. It allows systems to exchange data and information using compatible meanings and definitions. This layer of interoperability ensures that data exchanged between systems has the same meaning and context. For example, two systems may use different terminologies or vocabularies to describe the same concept. Semantic interoperability ensures that data is processed correctly, taking into account the underlying semantics and business rules. This is achieved through the use of standardized medical terminologies, such as SNOMED-CT [8] and LOINC [7]. With ongoing efforts to improve and standardize health information exchange, interoperability and its foundation [9] have the potential to revolutionize the way healthcare providers and organizations share information and deliver care. Since healthcare providers are under increasing pressure to provide patient-centered care that is safe, effective and timely, interoperability is critical to accessing information and making informed decisions. This ability of heterogeneous systems to exchange and understand the received data without any ambiguity helps reduce the risk of medical errors as well as reducing the work of healthcare professionals, which leads to overall improved patient outcomes.

The prevention of disease outbreaks is an important part of eHealth's goals. Effective procedures for registration, notification and reporting on communicable diseases are critical to achieving this objective. Timely and accurate reporting of communicable diseases can help healthcare officials detect outbreaks early and respond quickly to prevent further spread of the disease. It can also aid in identifying infected individuals and isolating them, including their contacts, thereby preventing transmission. Furthermore, disease reporting data can be used to plan and allocate resources for disease prevention and control, while regular analysis of disease data can help identify trends and patterns, allowing healthcare officials to implement targeted prevention strategies.

The Bulgarian Ministry of Health sets out a comprehensive description of the rules and procedures for registration and reporting of communicable diseases in Ordinance No. 21 on the Procedure for Registration, Notification and Reporting on Communicable Diseases (suppl. No. 55/2022) [2]. The ordinance mandates health service providers to report the cases but does not address the problems with ineffective data exchange, delays and the diversity of information systems in health care domain. The current situation in Bulgarian healthcare lacks the necessary interoperability between the disparate medical information systems, which leads to unnecessary repetitive work, delays, incomplete data and errors. To effectively control and prevent communicable diseases in Bulgaria, there is a need for a comprehensive and interoperable information system that supports real-time data exchange, monitoring and reporting of disease outbreaks. The system must allow integration with other systems and alternative ways for information input and output, for example, alternative input of laboratory results through an XML file based on a defined XSD schema.

The objective of this paper is to implement the wide set of requirements of Ordinance No. 21 for registration, notifying and reporting of patient cases with infectious disease. In the process of modeling and developing a software solution, terminologies and nomenclatures related to standardization and normalization of health data are reviewed, emphasizing the models providing semantic interoperability in data exchange and more specifically those enabling the unambiguous exchange of diseases and conditions like the International Classification of Diseases (ICD) [10]. These are fundamental to the exchange of medical data between heterogeneous information systems, both at the national and cross-border levels, and are invariably present in solutions that successfully provide semantic interoperability.

The following tasks arise from the described objectives:

1. Resolve eHealth issues related to lack of interoperability and research best practices for modeling and sharing clinical data.

2. Examine the requirements of the Bulgarian Ministry of Health, in connection with the procedures for registration, notification and reporting on communicable diseases.

3. As a result of the above, a solution proposal in the form of a three-tier application.

After considering the objectives set, a model is proposed and a concept is implemented for effective clinical data exchange in the context of the requirements of Ordinance No. 21 on the procedures for registration, notification and reporting on communicable diseases.

This paper is structured in four parts. Section 1 defines goals and objectives, introduces basic concepts and defines the problem area and hence the motivation for the proposed solution. Section 2 consists of an overview of the methodology for the proposed solution, architecture, and database model, including the specified requirements of Ordinance No. 21 on the procedures for registration, notification and reporting on communicable diseases. Based on the research carried out and the detailed study of the requirements in the regulation, Section 3 describes a proposed solution in the form of a three-tier web application, comprehensively covering the requirements issued by the Bulgarian Ministry of Health. Section 4 draws a conclusion, summarizes the results of the completed tasks and presents future perspectives.

2. Methods

This Ordinance No. 21 on the Procedure for Registration, Notification and Reporting on Communicable Diseases, issued by the Bulgarian Ministry of Health, outlines the steps required to register, notify, and report individuals who have infectious diseases, those who have been in contact with them, and carriers of such diseases. It also specifies the list of infectious diseases that require mandatory registration, notification, and reporting, together with a comprehensive list of criteria for each disease and based on those specific criteria, the rules by which the communicable disease case level is determined.

The disease criteria can be categorized as clinical, laboratory and epidemiological.

Clinical criteria include common and practical signs and symptoms of the disease that alone or in combination represent a clear or suggestive clinical picture of the disease. Clinical criteria give a general description of the disease and do not necessarily indicate all the signs to make a specific clinical diagnosis.

Laboratory criteria consist of a list of the laboratory methods used to confirm the case. Usually only one of the specified tests is sufficient for laboratory confirmation. For some diseases, laboratory criteria should only indicate a probable case. It becomes clear that each communicable disease has its own rules for determining the classification of the case, which require complex business rules to be followed and implemented for the realization of a unified software solution.

Epidemiological criteria are considered to be met when an epidemic link can be established. During the incubation period, an epidemic relationship is present in certain cases for the specific disease, such as human-to-human transmission, animal-to-human transmission, exposure to a common source of infection, consumption of contaminated food or drinking water, exposure from the environment, including laboratory exposure.

The cases of each communicable disease in the ordinance can be classified as follows:

Possible. This is a case classified as "possible" for reporting purposes. It is usually a case in which the clinical criteria described in the case definition are present but for which there is no epidemiological or laboratory evidence of the disease in question. The possible case definition has high sensitivity and low specificity. It allows the detection of most cases, but some false positive cases will be included in this category.

Probable. This is a case classified as "probable" for reporting purposes. It is usually a case where the clinical criteria and an epidemic relationship as described in the case definition are present. Laboratory tests for probable cases are specified only for some diseases.

Confirmed. This is a case classified as "confirmed" for reporting purposes. These cases are laboratory-confirmed and may or may not meet the clinical criteria described in the case definition. The definition of a confirmed case has high specificity and low sensitivity, meaning most of the cases collected will be genuine, although some might be missed.

Bulgarian healthcare requires the mandatory registration, notification and reporting of a large number of communicable diseases (currently 67). Each of these diseases has different case type delineation requirements, which demand complex and hard to track and enforce business rules for categorizing the cases of infectious disease. In addition, there is a need for rapid and reliable communication between multiple health care actors in the communicable disease disclosure (Figure 1), as well as eliminating the need to repeatedly enter the same information, for example, in multiple information systems and in the communicable disease book. Reliable storage and automated retrieval of various statistics filtered by area, time intervals and a particular disease or category of diseases are also required. The information

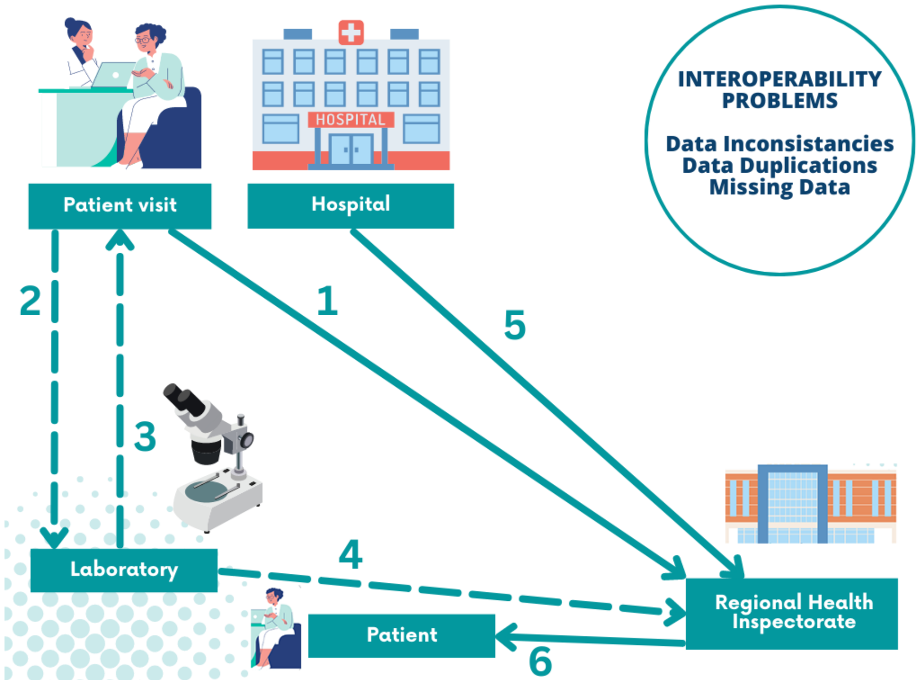

Figure 1. Data exchange model for registering a communicable disease

collected in this way can allow for more effective control in cases of quarantine and epidemics.

When registering and reporting an infectious disease in the Bulgarian healthcare system, participants typically follow a model for data exchange activities with the following sequence:

1. A patient visits his general physician (GP), the doctor evaluates the patient's condition and, provided the health state of the patient matches a rule for "possible", "probable" or "confirmed" statements for a communicable disease in the Ordinance, he creates an Express notification for the respective disease. This notification is sent for processing to the Regional Health Inspectorate (RHI). A direction for additional Laboratory tests is issued in the case of a "probable" or "confirmed" (Activities 1 and 2 in Figure 1).

2. In case the Laboratory exams confirm the communicable disease, the Laboratory creates an Express notification for the respective disease with a "Confirmed" statement both to the GP and the RHI office (Activities 3 and 4 in Figure 1).

3. When a communicable disease is established in a hospital, the Hospital creates an Express notification for the respective disease and sends it for processing to the Regional Health Inspectorate (Activity 5 in Figure 1).

The sequence of activities in Figure 1 demonstrates that the registration and reporting processes for communicable diseases in Bulgarian healthcare require multiple exchanges of information between various participants and heterogeneous systems.

A centralized information technology solution based on the three-tier architecture could significantly facilitate the complex processes related to the registration, exchange and storage of information as per the requirements of the ordinance. A three-tier architecture is a widely used model for developing applications that can provide several benefits for the procedures for registration, notification, and reporting on communicable diseases. This architecture separates the presentation layer, application logic layer, and data storage layer, which enables independent scalability, improves maintainability, and enhances security (Figure 2). Furthermore, the three-tier architecture facilitates fault tolerance, code reusability, and performance optimization, which can significantly enhance the overall efficiency and effectiveness of the system, including satisfying the possibility of integration with other systems. These benefits of the three-tier architecture make it a suitable choice for developing an information system for registration and reporting of communicable diseases, ensuring a robust and secure system that can handle the demands of public health emergencies.

In the context of developing an appropriate solution, choosing the appropriate type of database is a very important decision. Relational and non-relational databases have different strengths and weaknesses, and the choice depends on the requirements. Experiments have been performed with non-relational database but at this stage, a relational database is more suitable as most of the data is structured, well related and complex queries are required. This means that the data is stored in tables with predefined relationships between them. The database model behind the proposed solution consists of several tables that are interrelated through various types of relationships (Figure 3).

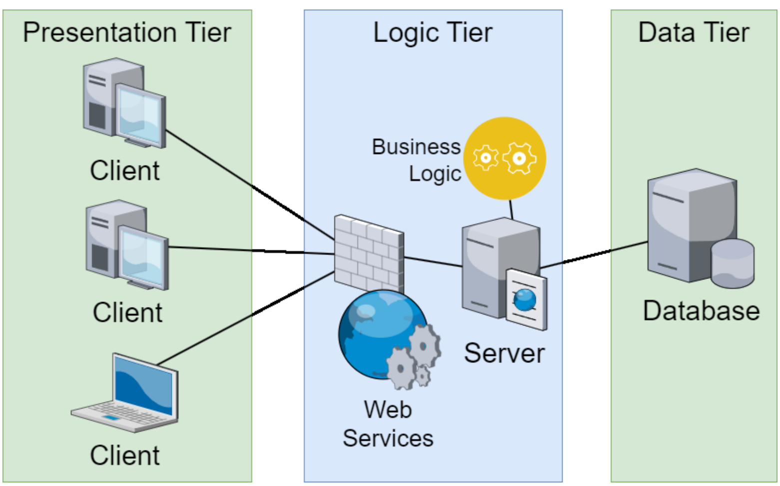

Figure 2. Architecture overview of the proposed information System

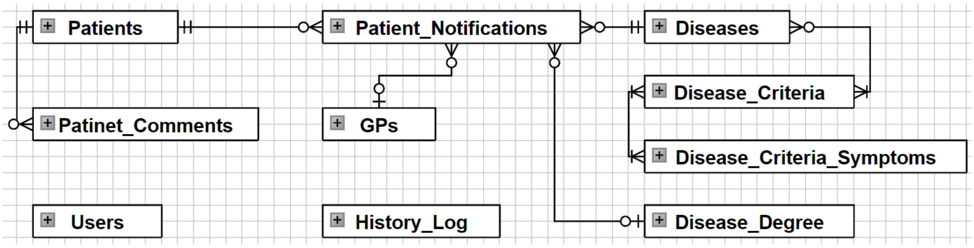

Figure 3. Relational database model view

Table Users contains the participants of the system, including managers, physicians, referrals and guests. Each user is assigned a unique identifier and access type. The table contains fields for user credentials, such as username and password, and additional fields to store user profile information.

Table Patients manages patient-related data, such as patient names, addresses, contact information, and other patient-specific entries. Each patient is assigned a unique identifier, and the table contains fields for demographic information, medical history, and other relevant data.

Table Patient_Comments is auxiliary and is used to store additional data related to the patient. It is linked to the Patients table in a many-to-one relationship, as each patient may have zero or multiple comments associated with them, where no comment could exist without an assigned patient to it. The table contains fields for comment text, date, author, and edit history.

Table Patient_Notifications is central table as it manages the disease notification data filled out by the physician for a specified patient and specified disease. The table contains multiple fields related to the procedure for registering communicable disease, including data for criteria symptoms such as clinical, laboratory, and epidemiological. Each notification is assigned a unique identifier and the table is linked to the Patients table in a many-to-one relationship, as each patient may have zero or multiple notification where no notifications could exist without an assigned patient to it.

Table GPs contains information about the general practitioner of the patient. It is linked to the Patient_Notifications table in a one-to-many relationship, as each notification may be associated with a specific general practitioner, where a GP may be linked with zero or multiple notifications. The table is not linked with Patients, as a patient is free to change its general practitioner. The table contains fields for general practitioners names, unique identifier numbers, addresses, contact information, and other relevant data.

Table Diseases contains a list and properties of each communicable disease defined by the Bulgarian Ministry of Health through Ordinance No. 21. It is linked to the Patient_Notifications table in a one-to-many relationship, as each notification is necessarily associated with a specific disease and a disease can be associated with multiple notifications. The table contains fields describing the disease name and its properties, symptoms, quarantine, and other relevant data.

Table `Disease_Criteria` stores specific criteria for each disease. The table contains fields for criteria name, description, and other criteria-relevant data.

Table `Disease_Criteria_Symptoms` contains a comprehensive list of all symptoms defined in the ordinance, such as diagnostic tests, symptoms, and other clinical indicators.

Table `Disease_Degree` is auxiliary and contains the type of communicable disease case, as determined by the ordinance. It contains fields for type names and descriptions. This information is stored in the database to ensure that the system is up-to-date with the latest regulatory requirements and to support future changes in the ordinance through the system interface without the intervention of programmers and need for code changes.

The applied cardinality restrictions on the relations of the described above tables are essential to ensuring compliance with business logic and rules. Cardinality restrictions define the types of relationships that can exist between entities in a database and specify the number of instances of one entity that can be associated with another entity. By enforcing cardinality restrictions, the database management system ensures that the relationships between entities are well-defined and adhere to the business logic of the system. This not only ensures the accuracy and consistency of the data but also helps to prevent errors and inconsistencies in the system.

3. Results

On the basis of the ordinance examined and best practices and standards in clinical data exchange, it becomes clear that an Information System for Registration and Reporting of Communicable Diseases (ISRRCD) in the form of a three-tier web application with a centralized database can provide several benefits in addressing the lack of interoperability between healthcare actors in relation to the process of registration of communicable diseases. These benefits include:

Improved efficiency. Facilitate the exchange of data between healthcare actors, reducing the need for manual and duplicated data entry and increasing efficiency. This can help healthcare providers save time and resources, enabling them to focus on patient care.

Enhanced accuracy. Ensure that the data entered is accurate and up-to-date. This can help improve the accuracy of disease tracking and reporting, and reduce errors in the reporting process.

Better communication. The three-tier web application can provide a platform for better communication between healthcare actors. The system can facilitate real-time data sharing, making it easier for healthcare providers to collaborate and coordinate efforts in addressing disease outbreaks.

Improved disease surveillance. Improve disease surveillance efforts by providing a comprehensive view of disease incidence and prevalence. This can help healthcare providers identify trends and patterns, enabling them to take appropriate actions to address disease outbreaks.

Enhanced public health outcomes. The solution can help improve public health outcomes by providing better data for public health decision-making. The system can help identify high-risk populations and facilitate the development of targeted interventions to reduce disease transmission.

Increased healthcare cost savings. Reduce healthcare costs by improving efficiency, reducing the need for manual data entry and duplicate testing, and enabling early detection and prompt treatment of communicable diseases.

The actors in the use case are as follows:

Managers. Responsible for managing the information that the physicians will be working with. This includes overseeing data entry, monitoring data quality, generating reports, and analyzing data to make informed decisions. They have full access to the system and are able to perform a wide range of tasks, such as creating and managing user accounts, defining data entry requirements, setting up workflows and approval processes, and monitoring system usage.

Physicians. Responsible for registering and reporting the communicable disease cases, updating patient records, and managing patient information within the system. Physicians have access to the application tier of the software, which includes the generated graphical interface based on a selected disease, logic and business rules that govern data entry and validation. Physicians are able to view and edit patient information as needed and are able to generate reports and other outputs related to patient care.

Referrals/Guests. The third category of participants have read-only access to the system, which means they are able to view patient records and other data but can not edit or modify any information. This access level is intended for external stakeholders, such as referral partners or external consultants, who need to view patient data for analysis or decision-making purposes but do not need to interact with the system on a daily basis.

The following use case for managing communicable disease case can be considered based on the actors described above:

A user of type *Manager* (a Regional Health Inspectorate specialist) logs into the system using their user account credentials. The manager can create and administer user accounts for the other participants and set access levels. Main task for the manager is to define the complete set of communicable diseases in Ordinance No. 21, which may be subject to change over time (adding new diseases), including the definition of symptoms and symptom criteria and the rules for each disease that define the case (possible, probable or confirmed). The Manager can also monitor system usage and generate reports to analyze and extract data.

A user of type *Physician* that can log in to the system using their user account credentials. The physician can register a communicable disease patient case, based on automatically generated graphical user interface for a selected disease. Furthermore, the Physician can view and edit patient information as needed to ensure data accuracy. The Physician also has access to generating reports and other outputs related to patient cases, such as lab results or treatment plans.

A user of type *Referral/Guest* can also access the system using their user account credentials. These users can view patient records and other data for analysis or decision-making purposes but cannot edit or modify any information in the system.

The following modular functionalities can be highlighted in the proposed information system for registration and reporting of communicable diseases:

Organizational module

This module is an essential component of the system and is designed to manage work-related data, particularly the data in the context of disease management. This module is responsible for generating and managing the necessary data required for the purposes of registering and reporting communicable disease cases.

Disease management. One of the key responsibilities of the organizational module is to manage disease-related data (Figure 4). This includes collecting and storing information for all diseases defined in Ordinance No. 21. This information is essential for the required processes for registration, notification and reporting.

Criteria and symptom management. The criteria and symptom management allow users to define the specific data required for the diagnosis and management of different diseases. The criteria can include various symptoms, laboratory results, or other factors that help determine the severity of a particular case. By defining these criteria, users can ensure that the system accurately assesses each case based on the relevant factors.

Management of the rules for case definition. Involves defining the specific rules that the system must use for determining the degree of a particular disease case. For example, the system may use a checklist of specific symptoms and laboratory results to determine whether a case is possible, probable, or confirmed. By defining these rules, users can ensure that the system accurately assesses each case based on the specific criteria defined for each disease. This process is automated so that the case is classified as soon as the registration is completed by the physician, and subsequently, the information and the case classification are automatically sent to the relevant participants in the health chain.

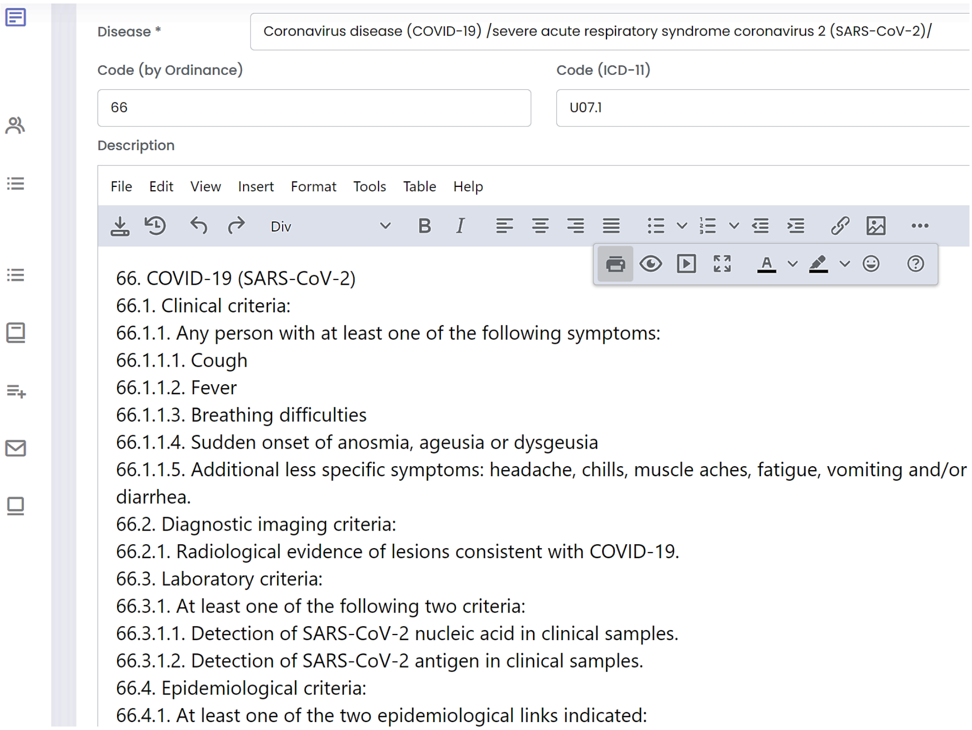

Figure 4. ISRRCD organizational module – overview of disease management

Physician's module

This module is a key component of the software system and is responsible for the registration and reporting of new communicable disease patient cases. The physician selects the specific disease from a pre-defined list of diseases. Once selected, the system automatically generates a graphical interface that includes a list of symptom criteria specific to the selected disease (Figure 5). This ensures that all relevant information is captured during the registration process. The Organizational module plays a key role in the Physician's module, as it defines the rules and symptom criteria for each disease. Based on these rules, the system automatically determines the degree of the case during the registration process. This ensures that cases are accurately classified as either possible, probable, or confirmed based on the specific symptom criteria for each disease described in Ordinance No. 21. Once the physician has completed the disease registration steps, the system automatically sends notifications to relevant parties (Figure 6), such as the regional health inspection, the Communicable Disease Book, the patient's GP, and the patient themselves (if their email is present). This helps to streamline communication and ensure that all relevant parties are informed of new cases without the need for manual data entry or notification. Export of the registration data to an XML file is provided so that it

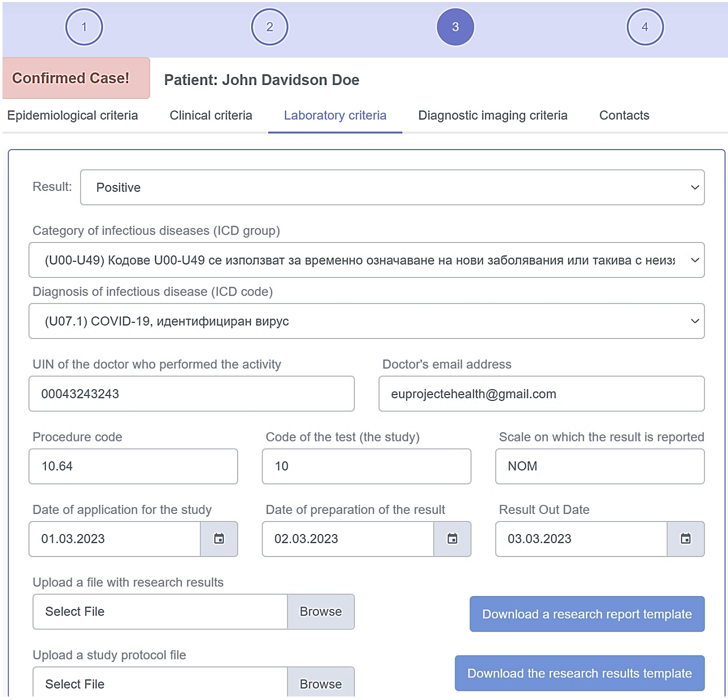

Figure 5. ISRRCD physician's module – registering a case of a communicable disease (from XML file)

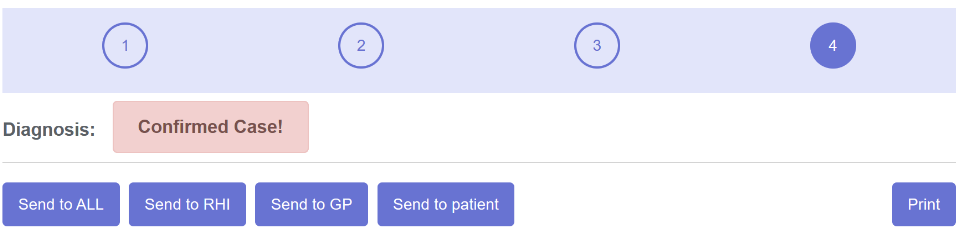

Figure 6. ISRRCD physician's module – registering a case of a communicable disease (step 4)

can be transmitted to other systems, thus achieving integration with external systems. The XML file schema might be subject to change according to the specific requirements of the foreign system.

Statistical module

Statistical data is important in preventing disease outbreaks as it helps identify patterns and trends of disease occurrence, track the effectiveness of prevention measures, and predict future outbreaks. This information allows public health officials to make evidence-based decisions and take proactive measures to protect public health. Therefore, the statistical module is an essential part of the proposed software system, which allows participants, including Managers, Physicians and Referrals/Guests to gain valuable insights into communicable disease incidence and prevalence within a specific area and time period. By using various filters like time intervals, location, and a specified disease or group of diseases, the module generates statistical reports and visualizations that help to identify trends, patterns, and correlations in the data.

One of the key components of the statistical module is its ability to produce visualizations that highlight critical information (Figure 7). Users can view data in different formats, such as tables, graphs, and charts, which makes it easier to identify areas of high disease incidence, track disease outbreaks, and evaluate the effectiveness of prevention and control measures. For example, a heat map can show the geographical distribution of cases, which can assist the specialists in identifying regions with high disease incidence.

The statistical module supports various statistical analyses that can help the specialists recognize trends and patterns in the data, predict future disease incidence, and aid in decision-making around prevention, control, and treatment. Public health officials, epidemiologists, and other stakeholders involved in disease surveillance, management, and research can utilize the features of the module to gain valuable insights into disease incidence and prevalence. The module is user-friendly and accessible, allowing the easy generation of reports and visualizations.

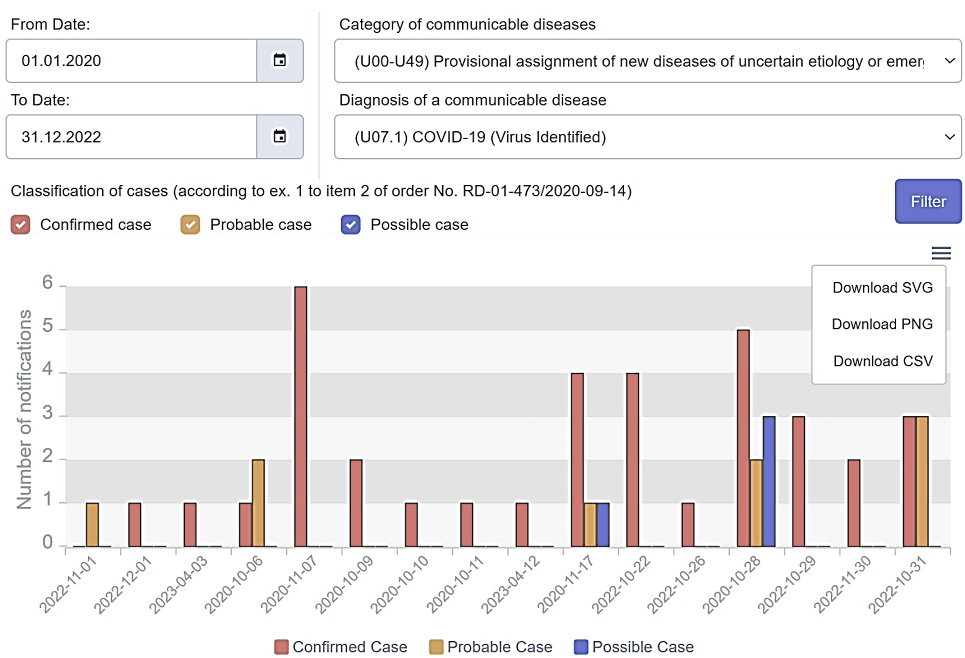

Figure 7. ISRRCD – statistical module view

Furthermore, the statistics module can provide valuable insights into the effectiveness of public health policies and interventions. By tracking disease incidence over time, healthcare providers and policymakers can evaluate the impact of different interventions and adjust their approach as necessary. This can lead to improved health outcomes for patients and reduced healthcare costs, as well as more effective public health policies and interventions. Ultimately, this can contribute to improved population health and well-being.

4. Conclusion

Based on the requirements in Bulgaria related to the registration and reporting of communicable diseases, reviewing the current situation in data exchange and after an extensive literature review, it becomes clear that solving the problems related to the lack of interoperability in Bulgarian eHealth is essential in the fight against communicable diseases, where quick and efficient information exchange is necessary to contain the spread of the disease.

The benefits of a centralized information system, which provides semantic interoperability and integration with other systems, are significant in contributing to improved public health outcomes by enabling earlier detection and more targeted interventions. The ability to monitor the epidemiological situation in real-time can help reduce the spread of disease and ultimately lead to better health outcomes for patients. The centralized system also allows for a more coordinated and efficient use of resources, reducing healthcare costs and increasing overall efficiency.

In conclusion, a well-designed information system for registration and reporting of communicable diseases with automated data exchange and statistical modules can make a significant contribution to public health by improving disease control and prevention efforts. It is therefore important to invest in such system to ensure that healthcare providers and policymakers have the tools they need to make informed decisions and effectively respond to disease outbreaks.

In the effort to solve these problems, the following achieved tasks can be summarized:

- A prototype of an Information System for Registration and Reporting of Communicable Diseases (ISRRCD) under Ordinance No. 21, issued by the Bulgarian Ministry of Health, has been modeled, developed and tested.

- Automated application of rules for classification of communicable disease cases.

- Registration of communicable disease cases in the Register of Infectious Patients in the Regional Health Inspectorate has been automated.

- The exchange of messages between participants in the process of registering communicable disease cases is automated, saving the time needed for data re-entry and thereby reducing errors.

- Visualization of statistics with summary data for the communicable disease cases stored in the database, preparation of reports by time interval, region, disease and other signs have been implemented.

The proposed software has been tested with real data and consultations have been carried out with specialists from the regional health inspectorates. A video demonstration of the current functionalities can be found on [1]. The software is at the prototype stage and the following directions for future work can be summarized: Further integration with other systems – alternative ways for entering data are provided, for example, through XML files according to a defined XSD scheme. These integrations are of significant importance in reducing the time required for duplicate data entry as well as reducing the possibility of errors in the process of integrating heterogeneous systems, for example, integration with various laboratory information systems.

The developed system can be transformed into a cloud-based data management platform, which will further eliminate human intervention in data entry and management when reporting cases of communicable disease patients. In this case, further research would be necessary in relation to requirements for the protection of personal data.

Acknowledgements

This work is supported by the Fund for Scientific Research at Sofia University "St. Kliment Ohridski" under Grant 80-10-8/11.04.2023.

References

[1] S. Abanos, Information System for Registering, Notifying and Reporting on Communicable Diseases, 2023. [Online]. Available: https://youtu.be/XBz0NcJqIhM [Accessed 08 2023].

[2] Bulgarian Ministry of Health, Ordinance No. 21 on the Procedure for Registration, Communication and Reporting on Communicable Diseases, 2022 (in Bulgarian). [Online]. Available: https://www.mh.government.bg/media/filer_public/2015/04/17/naredba-21-ot-2005g-spisak-zarazni-bolesti-red-registratsia.pdf [Accessed 08 2023].

[3] Health Resources & Services Administration (HRSA), What is telehealth?, 29 06 2022. [Online]. Available: https://telehealth.hhs.gov/patients/understanding-telehealth [Accessed 08 2023].

[4] Healthcare Information and Management Systems Society, Interoperability in Healthcare, 2023. [Online]. Available: https://www.himss.org/resources/interoperability-healthcare [Accessed 06 2023].

[5] HealthIT.gov, What is an electronic health record (EHR)?, 2022. [Online]. Available: https://www.healthit.gov/faq/what-electronic-health-record-ehr [Accessed 05 2023].

[6] Pan American Health Organization (PAHO) and World Health Organization (WHO), Introduction to Semantic Interoperability, 2021. [Online]. Available: https://iris.paho.org/bitstream/handle/10665.2/55417/PAHOEIHIS21023_eng.pdf [Accessed 04 2023].

[7] Regenstrief Institute, Logical Observation Identifiers Names and Codes, 1994. [Online]. Available: https://loinc.org/ [Accessed 07 2023].
[8] SNOMED International, SNOMED International: Leading healthcare terminology, worldwide, 2023, [Online]. Available: https://www.snomed.org/ [Accessed 06 2023].
[9] Wemedoo – Clinical Information Specialists, SNOMED CT, LOINC, and ICD-10 – The Foundation of Semantic Interoperability, 2021. [Online]. Available: https://wemedoo.com/snomed-ct-loinc-and-icd-10/ [Accessed 04 2023].
[10] World Health Organization, International Statistical Classification of Diseases and Related Health Problems (ICD), 2019. [Online]. Available: https://www.who.int/standards/classifications/classification-of-diseases [Accessed 07 2023].
[11] World Health Organization, WHO E-health, 2022. [Online]. Available: https://www.emro.who.int/health-topics/ehealth/ [Accessed 02 2023].

Received on April 13, 2023
Accepted on May 7, 2023

Simeon Abanos, Evgeniy Krastev and Petko Kovachev

Faculty of Mathematics and Informatics
Sofia University "St. Kliment Ohridski"
5 James Bourchier Blvd.
1164 Sofia
BULGARIA

E-mails: simeonabanos@gmail.com
eck@fmi.uni-sofia.bg
az@petko.info

Dimitar Tcharaktchiev

Medical University of Sofia
University Hospital of Endocrinology
2, Zdrave Str
1431 Sofia
BULGARIA

E-mail: dimitardt@gmail.com

ГОДИШНИК НА СОФИЙСКИЯ УНИВЕРСИТЕТ „СВ. КЛИМЕНТ ОХРИДСКИ“
ФАКУЛТЕТ ПО МАТЕМАТИКА И ИНФОРМАТИКА
Том 110

ANNUAL OF SOFIA UNIVERSITY “ST. KLIMENT OHRIDSKI”
FACULTY OF MATHEMATICS AND INFORMATICS
Volume 110

ENTERPRISE DATA AND SEMANTIC MODELLING: CONCEPTUAL MODEL OF INFORMATION TECHNOLOGY INCIDENT MANAGEMENT

KRISTINA ARNAOUDOVA and MARIA NISHEVA-PAVLOVA

Knowledge management methods and their efficient implementation across the organization determine sound and resilient management of processes. This paper studies the semantic integration of enterprise data sources essential to service management processes. Implementing a semantic layer within the enterprise architecture uses various tools, methods, and techniques. The semantic conceptual model unifies and implements intelligent integration of multiple data sources across the enterprise, achieving consistency and more accessible interpretation. Specifically, we draw our attention to incident and problem management within enterprises. We propose an ontology – a conceptual model for the incident management process. The incident ontology presented as an intelligent data integration layer component aims to achieve operational excellence. Besides, this ontology is a fundamental part of the proactive process in problem management. An ontology as a logic-based system supports integrity validation. It infers new, no explicitly modeled facts in the problem domain, thus helping experts better analyze and understand the problem. We discuss the conducted experiment results with the proposed in this article conceptual model using the enterprise knowledge graph platform. It can be perceived as a framework for a query-answering system with components, including ontology schema, data mapping, and classification methods for data graph enrichment.

Keywords: knowledge management, knowledge representation, ontology, semantic model, hybrid classification model, enterprise semantic layer, enterprise knowledge graph, incident management, ITIL, problem management

CCS Concepts:

- Information Systems~Information retrieval~Document representation~Ontologies;
- Information Systems~Information retrieval~Retrieval models and ranking;
- Information Systems~Information retrieval~Evaluation of retrieval results

DOI: 10.60063/GSU.FMI.110.25-36

1. Introduction

The business processes within an organization are of great variety. Their automation includes many applications. The enterprises are targeting improvement in many aspects, including better customer experience and manageable processes, where data integration is one of the main challenges. There are many approaches, platforms, and tools. With the significant increase in data volume, the complexity of the analysis and insight discoveries grows considerably, becoming an ultimate challenge. The centralized data store is one of the often-used components in the overall enterprise architecture. In many aspects, it is a costly approach that does not give results as expected. Vast methods and platforms are addressing this topic. Some ensure centralized storage and consistent reporting within the organization but are designed to solve specific tasks and reflect various aspects of applications, but not the business domain and the underlying relationships. Thus, it does not represent the problem domain and the concept's relations. A complementary approach is an intelligent integration using a virtual unified data layer.

The type of data assets in organizations, structured or unstructured, increases the difficulties in various tasks of the integration processes. The text data assets are a source of intensive knowledge. They are the primary source of additional information for the methods that might add other steps in the pipeline, like necessary preprocessing steps or concept mapping, and affirm challenges before information extraction. The lack of unified metadata or catalog leads to ambiguity and inconsistent automation. This paper presents a conceptual model essential for semantic data integration and apprehends the metadata for the data sources across one organization, allowing an ontology-based data access approach [18].

The particular domain is Information Technology Service Management (ITSM). The IT Infrastructure Library (ITIL) is an IT service management framework that outlines best practices for delivering IT services [3, 17]. Implementing a good ITSM process as ITIL has numerous benefits for an organization [6]. The complementary component to the processes is the semantic layer – a technique to implement intelligent integration and interoperability. The specific process we are reviewing is Information Technology service incident management, a part of the processes and practices of Information Technology Infrastructure Library (ITIL) methodology, widely adopted in many enterprises. An incident and its resolution are critical for service reliability and organization credibility. Extracting and enriching organizational knowledge is a fundamental part of the pipeline of construction of our holistic view of corporate IT assets. Our model, an essential component in the semantic data layer, may contribute to operational excellence proactively.

2. Conceptual model

IT service management encompasses many practices, including the operational one of resolving incidents. According to ITIL Operation management, the incident is [3]: "An unplanned interruption to an IT Service or reduction in the quality of an IT Service. Failure of a configuration item that has not yet impacted the service configuration item is also an incident, for example, failure of one disk from the mirror. Incident management deals with all incidents; this can include failures, questions, or queries reported by the user by technical staff or automatically reported by the event".

Solving an incident that has interrupted service is critical and includes determining the failure, taking the right action, and restoring the service's regular work. Implementing a fast process of resolving incidents is an indispensable asset of an organization where the critical aspect is to include all possible sources of information. Gathering expertise about the incident is the expert responsibility, defining the incident, including essential attributes and a text description. The enrichment using other, typically from unstructured data, information extraction techniques could be valuable for more accurate first-level incident analysis.

2.1. Conceptual schema. Knowledge and data

The incident management process can benefit from semantically modeled data in finding new information, streamlining the resolution in a resilient and predictive way. The classical approach for enrichment is to infer new implicit facts by applying logical axioms embedded in the schema. Automatically deduce a taxonomy with subsumed classes with our categories or infer new relations among them. The solution can be found using expert knowledge or knowledge from already solved incidents stored in the database.

The priority assessment is among many steps in incident management, essential for the escalation process. An integral part of the process is the estimation of impact and urgency. The impact reflects how many configuration items such as storage, servers, or applications are affected, and the urgency reproduces how fast other configuration items will be affected.

Our proposed conceptual schema aims to categorize the incident according to different characteristics automatically. For example, the incident priority, impact, and urgency can be defined with appropriate axioms as a part of the conceptual schema. We can categorize according to business rules and components and consequently have better ideas that might complement determining possible causes.

A fundamental component of the data integration approach in the semantic data layer is the conceptual model, realized further as ontology. The proposed article may help implement a non-incident-centric strategy for operational excellence following the proactive understanding of the incident management process based on a semantic data graph. We have planned various experiments with the proposed model. We have used a knowledge graph platform as a design schema integration data tool, query answering, and database for them. The knowledge platforms are integration tools that combine different approaches for finding insights and have functionalities for semantical enrichments. They can link various data sources that could be extremely helpful in resolving incidents, including regulations and organization procedures.

Semantic network

Different descriptive formalisms for knowledge representation are organized as structural and formal languages. The Semantic Web is perceived as a semantic network of immense scale. The semantic networks as a knowledge representation structured formalism are at the heart of the described model. The semantic network is a structured formalism, and it is many variants John Sowa defines [2]: "A semantic network or net is a graph structure representing knowledge in patterns of interconnected nodes and arcs. Computer implementations of semantic networks were first developed for artificial intelligence and machine translation, but earlier versions have long been used in philosophy, psychology, and linguistics." All semantic networks are graphically represented knowledge that supports automated systems for reasoning. Some versions are informal, but others are formally defined logic systems [2].

Ontology

The concepts and the relations among them should be unambiguously defined, and that is the main target of ontology as metadata about domain entities and relationships among them. Tom Gruber establishes the idea of ontology as [13] "an explicit specification of a conceptualization", creating machine-readable explicit formalization of a domain. An ontology consists of concepts and relationships constrained by domain-specific business rules.

The knowledge is represented with formal languages; those based on first-order predicate calculus are RDF Schema (RDFS) [10] for concept, relations, and restriction, and the Ontology Web Language (OWL) [8] as an RDFS extension. The fundamental standard Resource Description Framework (RDF) [11] widely adopted a graph data model of objects and its presentation in subject-predicate-object triples. Applying the Shape Conceptual Language (SHACL) [12] enables distributed validation over data silos and rules described by OWL 2 RL [7] and SWRL [16].

OWL is probably the most used knowledge representation language. The OWL sublanguages OWL Lite, DL, and Full enable the knowledge engineer to balance expressiveness and semantic capabilities while ensuring efficient inference.

SHACL is an RDF-based language that can check for integrity constraints not globally as OWL 2 axioms. Instead, it can be performed for each data set. SHACL is appropriate for expressing data integrity, while the traditional data integrity constraints are global and thus not readily adopted to distributed data.

Simple Protocol and RDF (Resource Description Framework) Query Language SPARQL [14] can define queries across diverse data sources, whether the data is stored natively as RDF or viewed as RDF. The result of a SPARQL query can be a data set or an RDF graph.

The above techniques are used in various realizations for enterprise-wide usage, ensuring interoperability and state-of-the-art enterprise management processes.

2.2. Conceptual model

The ontology concepts, relations, and constraints are central to the proposed semantic model. As such, the conceptual schema design results from concentrated domain knowledge analysis. Our model follows the top-down approach based on methodology and human expertise. The high-level concepts and their relations are presented in Figure 1.

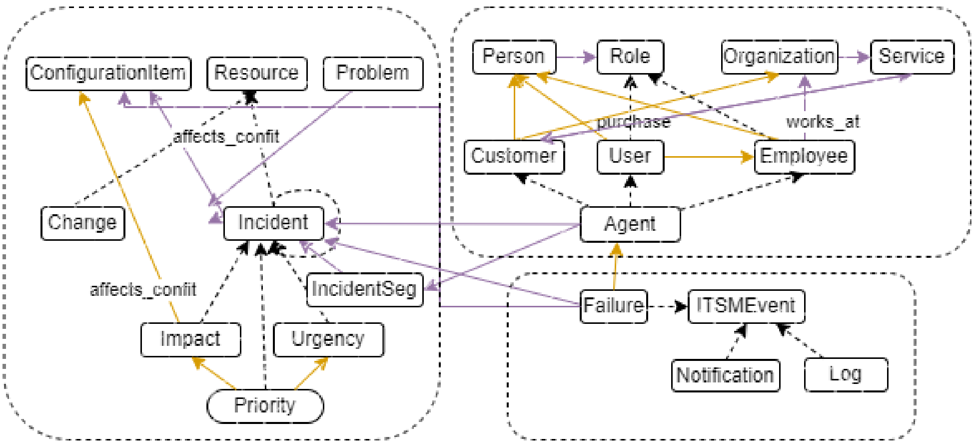

Figure 1. Incident conceptual model

2.2.1. Concepts

The main concepts, designed as owl classes, are *Request*, *Incident*, *Agent*, *Resource*, and *Configuration Item* examples: server, storage, application, service, and agents. The request for Resources could be for implementing a change, incident resolution, or service desk request. An example of an *Application as a subclass of Resources* is associated with its modules and architecture. An Agent is specialized as a person, and a person has different roles. The important roles in the incident process are client, employee, and user.

2.2.2. Ontology specification

OWL Class: inc:Incident
The class *Incident* includes different types of service interruption with the attributes describing the object.
Sub-class-of: inc:Request
Restriction: domain of object property comprise_seg

OWL Class: inc:Agent
The class *Agent* groups the agents or system who notifies of the service interruption.
Sub-class-of: owl:Thing
Restriction: domain of object property ann_incident

OWL Class: inc:Customer
The class *customer* includes those person or organization entities who have acquired a service or product from an organization.
Sub-class-of: owl:Role
Restriction: domain of object property *purchases*

OWL Class: inc:User
The user denotes persons who work at an organization or customers of an organization and are authorized to use the services of the organization.
Sub-class-of: owl:Role
Restriction: domain of object property register

OWL Class: inc:ISMEvent
The class *ITSMEvent* includes the system operation messages
sub-class-of: owl:Thing

OWL Class: inc:Role
The class *Role* includes different roles that a person or organization may have
sub-class-of: owl:Thing

OWL Class: inc:ConfigurationItem
The class *Configuration item* includes a comprehensive type of IT assets
sub-class-of: owl:Thing
Restriction: domain of object property subject_of

3. Enterprise knowledge graph

"Knowledge graph" is a term proposed by Google in 2012 as RDF graph data model, without details about its realization [18]. The knowledge graph may be described as a framework for realizing a data layer for querying information across various data sources, thus unifying different silos and data access across many sources. It may be implemented as a database or virtually with a transformed query. Its main component as a knowledge-based system is the data schema or ontology. It has embedded logic axioms, which could be formalized differently, performing constraints and inferring additional, not explicitly modeled concepts or relations. Statistical reasoning is another approach for enriching the knowledge graph in many knowledge platforms.

The pipeline is not fully automated but could be partially realized and enriched using semantic parsing, natural language understanding, machine learning, and logic inference. Different profiles ensuring different levels of expressiveness are adopted to manage the performance limitation of the reasoning engines. The semantic layer data's graph unifies distributed silos. The construction of a knowledge graph is supposed to have good scalability and performance, which is achieved with an on-demand inference or approximate one for knowledge graph storage virtualization and querying.

To validate the proposed conceptual mode for incident management, we have defined the ontology using OWL and the mapping procedure for conceptualizing the data. The processed experiments with the knowledge platform Stardog [15] based on a graph model that supports various formats and reasoning, including RDFS, OWL, SHACL, and SWRL. Following are some excerpts from the schema and screen with reasoning.

3.1. Graph data

We use data from an open data repository – the Incident management process enriched event log Data Set[1]. The dataset reflects information about incidents and the workflow for incident resolution, related agents, affected configuration items, and the different stages of the process. It is an event log of an incident management process extracted from data gathered from an instance of the ServiceNow™ platform. The event log is enriched with data loaded from a relational database underlying a corresponding process-aware information system. Information is anonymized for privacy. The instances are 141 712 events, but 24 918 incidents. The concept mapping is created by file with platform-specific formats associating the data row to the classes and relationships. The result is confirmed by the created incident nodes with attributes, configuration items, and other concepts.

3.2. Incident data graph

The primary type of objects are concepts, designed as owl classes, and are *Request*, *Incident*, *Agent*, *Resource*, and *Configuration Item.* The incident is instantiated from the customer, user, or system, and an incident announcement is an event using a specific component. We model the incidents as sequences of segments denoting the different steps in the incident resolution ontology, following the design patterns for ordering relationships [1]. For example, the introduction of the incident relates to opening the next step could be the investigation performed by the assigned expert, then closing the incident.

3.2.1. Priority and impact

The incident priority is formulated as a defined class, categorizing the incident in a subcategory, i.e., high-priority-incident. The incident is classified according to other parameters like business service affected, urgency, or impact. The priority formulation is in terms of impact and urgency, and the incident impact is defined in terms of affected configuration items. For example, the classification axioms infer a medium impact if the affected items are more than substantial value, which is specific to the organization.

[1] https://archive.ics.uci.edu/ml/datasets/Incident+management+process+enriched+event+log

3.3. Configuration item data graph

The location is identified as a physical place where a person or organization may reside or a virtual address. The configuration items in Figure 2 and Figure 3 are the operations assets and the configuration management database (CMDB) entities with the corresponding hardware, application, and services subclasses. The configuration items are the range of the affected configuration item relation.

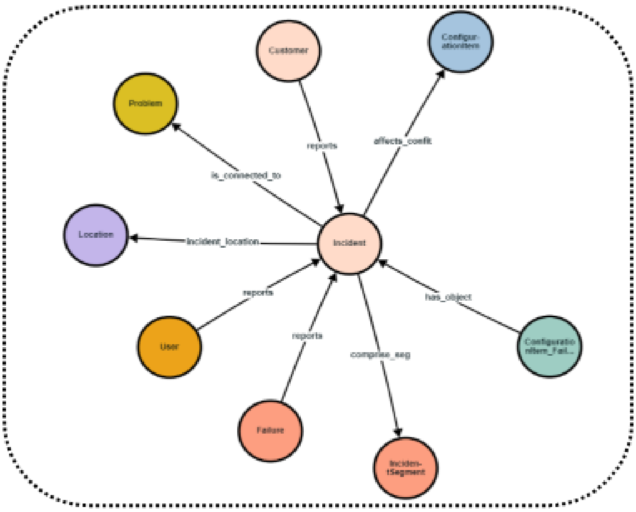

Figure 2. Incident schema

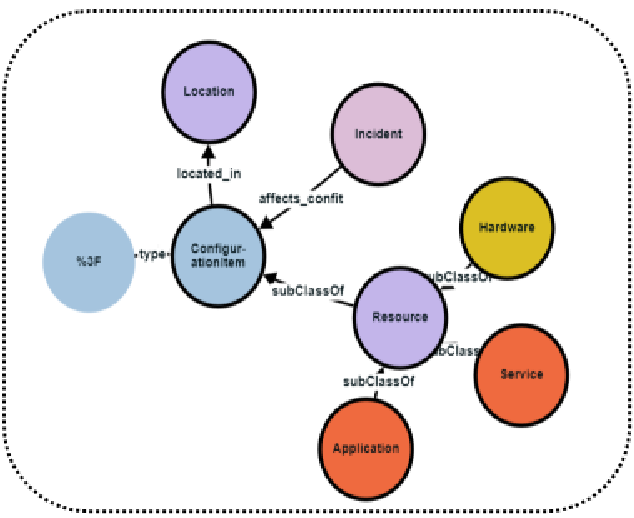

Figure 3. Configuration item type and instances

3.4. Agent data graph

The *Agent* class is specified in the subclasses *person*, *customer*, and *system notification*. The concept underpins the incident logging step of the incident management process. The incident management process follows several steps, among them incident categorization, prioritization, assignment, and resolution.

Organization

The *organization* class in Figure 3 relates to the *owner* and *representative*, which are the specification of class roles. An organization instance has employees and an internal structure defined in an organizational hierarchy in employee groups. The *organization* produces *products* that clients purchase.

3.5. Role schema

A person or organization may have different temporal roles, presented in Figure 4 and Figure 5, like a company's customer, user, owner, or representative. The constraint for a customer is to have at least one product. In Figure 5, we can follow the customer who has instantiated the presented incidents object.

ISTM event

Figure 6 illustrates the system's event interpretation – the notification of the status code of the systems, i.e., warning or failure. The failure of some components may automatically create an incident in the workflow. The typical channel is service support, where a customer reports the incident while users open an incident due to a malfunctioning component of the IT environment.

4. Classification

Our planned experiments include inferences based on the OWL 2 Lite and the OWL 2 DL [9] description logic reasoner, and the OWL 2 RL semantics with SWRL support. The instances are classified with eager reasoning based on axioms. In Figure 7, the experiments with the prioritization process assign a member to the related incident category – medium priority, according to its impact and urgency. It is realized with appropriate ontology axioms and rules. Such classification step could be included in the incident resolution pipeline as part of the solution for automatic inference or validation of an existing manual priority assessment process. Thus, we can derive new, non-explicitly defined facts.

Another approach to classification is reasoning, performed at query time. In the performed experiments, the data set was relatively limited. We plan various experiments over different data sources to reveal the model's capability further. Among them, virtual and materialized unification are planned. As a machine learning aspect, we plan to apply the inductive approach to predict values or search for

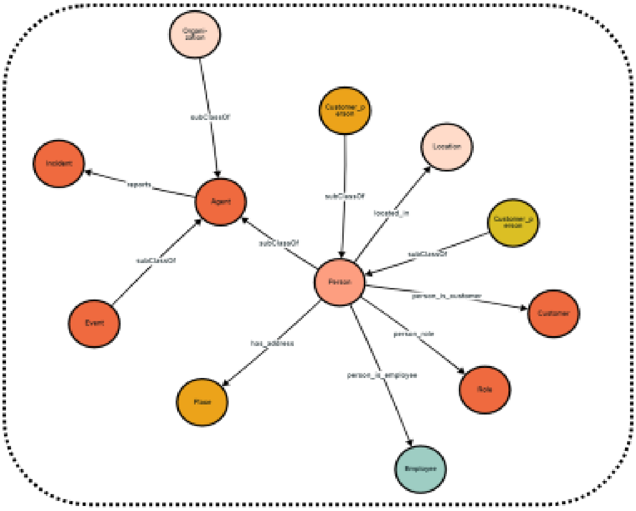

Figure 4. Agent, Person, and Role class

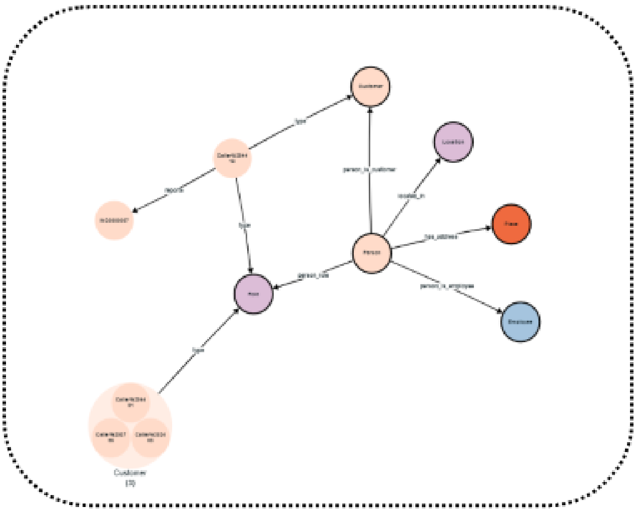

Figure 5. Role class and instances and relations

similarities in the data embedded in the graph [5]. Another intended experiment in the future is to use suitable NLP techniques to extract additional information about the incident from text data sources.

5. Conclusion

The semantic approach to search engines within an organization can significantly improve the processes within the organization, and the ontology approach

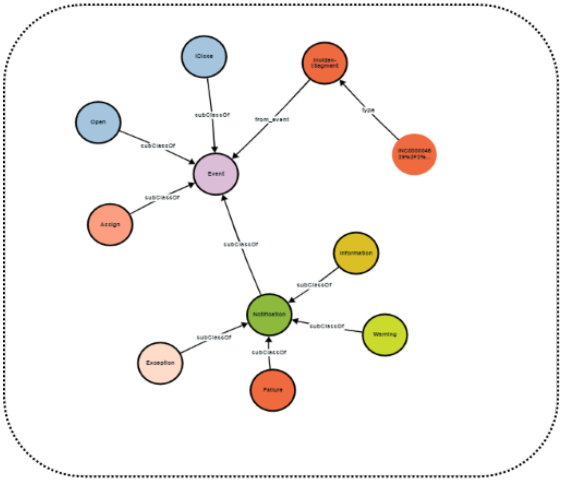

Figure 6. ITSM Event

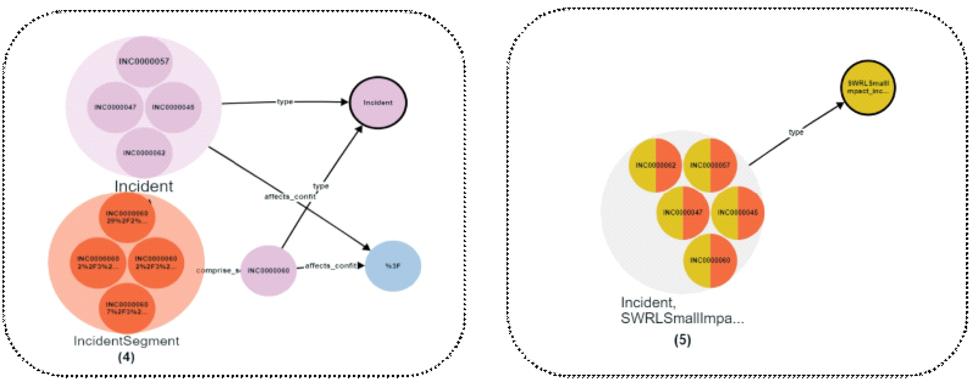

Figure 7. Incident classification

may solve different generic tasks within the organization [4]. It may rely on graph data, comprehended holistically, including various data silos. Constructing a semantic data model depends on multiple methods and techniques to build, support and enrich the enterprise's knowledge assets. Our experiments on the operational process for IT incident management with semantically modeled proactive incident discovery show promising results. Further development of such an approach can potentially leverage the organization's operational excellence.

References

[1] Defining N-ary Relations on the Semantic Web, https://www.w3.org/TR/swbp-n-aryRelations.

[2] T. Gruber, Principles for the design of ontologies used for knowledge sharing, Int. J. Hum. Comput. Stud. 43 (1993) 907–928.

[3] ITIL – IT Infrastructure Library, https://www.ibm.com/topics/it-infrastructure-library [Accessed on 11 April 2023].
[4] K. Kaloyanova, I. Patias and B. Robev, Prototyping data management model and tasks, in: CEUR Workshop Proc., ISSN 1613-0073, Vol. 3061 (2021) 256–264.
[5] K. Liang, L. Meng, M. Liu et al., A survey of knowledge graph reasoning on graph types: Static, dynamic, and multimodal, arXiv:2212.05767v6 [cs.AI].
[6] M. Marrone, L. Kolbe and M. Lutz, Impact of IT Service Management Frameworks on the IT Organization, Bus. Inf. Syst. Eng. **3**(1) (2011) 5–18, https://doi.org/10.1007/s12599-010-0141-5.
[7] OWL 2 RL in RIF (2nd Edition), https://www.w3.org/TR/rif-owl-rl/.
[8] OWL 2 Web Ontology Language. Structural Specification and Functional-Style Syntax (2nd Edition), W3C Recommendation, 11 December 2012, https://www.w3.org/TR/owl2-syntax/.
[9] OWL 2 Web Ontology Language Profiles, https://www.w3.org/TR/owl2-profiles/#Profile_Specification_3.
[10] RDF Schema 1.1, https://www.w3.org/TR/rdf-schema/.
[11] RDF 1.1 Concepts and Abstract Syntax, https://www.w3.org/TR/rdf11-concepts/.
[12] Shapes Constraint Language (SHACL), W3C Recommendation, 20 July 2017, https://www.w3.org/TR/shacl/#terminology.
[13] J. Sowa, Semantic Networks, in: Encyclopedia of Artificial Intelligence, ed. by S. Shapiro, 1987.
[14] SPARQL 1.1 Query Language, W3C Recommendation, 21 March 2013, https://www.w3.org/TR/sparql11-query/.
[15] Stardog Tutorial, https://docs.stardog.com/tutorials/.
[16] SWRL: A Semantic Web Rule Language Combining OWL and RuleML, https://www.w3.org/Submission/SWRL/.
[17] What is ITIL? Your guide to the IT Infrastructure Library | CIO, https://www.cio.com/article/272361/infrastructure-it-infrastructure-library-itil-definition-and-solutions.html.
[18] G. Xiao, D. Calvanese, R. Kontchakov et al., Ontology-based data access: A survey, in: Proc. 27th Int. Joint Conf. on Artificial Intelligence IJCAI 2018, Survey Track, 5511–5519, https://doi.org/10.24963/ijcai.2018/777.

Received on March 31, 2023
Accepted on May 7, 2023

Kristina Arnaoudova

Faculty of Mathematics and Informatics
Sofia University "St. Kliment Ohridski"
5 James Bourchier Blvd.
1164 Sofia
BULGARIA

E-mail: k.arnaoudova@fmi.uni-sofia.bg

Maria Nisheva-Pavlova

Faculty of Mathematics and Informatics
Sofia University "St. Kliment Ohridski"
5 James Bourchier Blvd.
1164 Sofia
BULGARIA

Institute of Mathematics and Informatics
Bulgarian Academy of Sciences
Acad. G. Bonchev Str., Bl. 8
1113 Sofia
BULGARIA

E-mail: marian@fmi.uni-sofia.bg

ГОДИШНИК НА СОФИЙСКИЯ УНИВЕРСИТЕТ „СВ. КЛИМЕНТ ОХРИДСКИ“
ФАКУЛТЕТ ПО МАТЕМАТИКА И ИНФОРМАТИКА
Том 110

ANNUAL OF SOFIA UNIVERSITY “ST. KLIMENT OHRIDSKI”
FACULTY OF MATHEMATICS AND INFORMATICS
Volume 110

AN IMPROVED BULGARIAN NATURAL LANGUAGE PROCESSING PIPELINE

MELANIA BERBATOVA and FILIP IVANOV

In this paper, we present a language pipeline for processing Bulgarian language data. The pipeline consists of the following steps: tokenization, sentence splitting, part-of-speech tagging, dependency parsing, named entity recognition, lemmatization, and word sense disambiguation. The first two components are based on rules and lists of words specific to the Bulgarian language, while the rest of the components use machine learning algorithms trained on universal dependency data and pretrained word vectors. The pipeline is implemented in the Python library spaCy (`https://spacy.io/`) and achieves significant results on all the included subtasks. The pipeline is open source and is available on Github (`https://github.com/melaniab/spacy-pipeline-bg/`) for use by researchers and developers for a variety of natural language processing and text analysis tasks.

Keywords: natural language processing, language pipeline, word sense disambiguation

CCS Concepts:

• Applied computing~Document management and text processing~Document capture~Document analysis

1. Introduction

A language pipeline consists of a sequence of steps targeted towards processing and analyzing natural language data. A typical language pipeline might include steps such as tokenization, part-of-speech tagging, parsing, and semantic analysis among others. These steps are used as a preprocessing stage in many different tasks and applications that involve analyzing human language.

Large number of language pipelines are built with the Python natural language processing library spaCy, which offers easy and flexible way to create such systems. A spaCy pipeline can combine predefined components, such as tokenizer and part

DOI: 10.60063/GSU.FMI.110.37-50

of speech tagger, as well as custom developed components and other functions, depending on the end goal of the system.

There are previous works for building a Bulgarian pipeline based on previous versions of spaCy [18] or custom-built software [19]. However, since these works have been published, new neural-based based algorithms for tasks such as lemmatization have been put into practice, which improve the performance and also facilitate the automatic evaluation.

Currently in spaCy v.3, there are trained pipelines for more than 20 languages, including low-resource languages from the Balkans, such as Macedonian[1], Greek [17], and Romanian, but not Bulgarian. A language pipeline with good performance is crucial both for conducting research in the field of word processing and other related areas and for creating software applications for various purposes.

The goals of the current work are:

- to create an end-to-end, open-source pipeline for the Bulgarian language in spaCy v.3;
- to improve available lists of tokenizer exceptions and stop words and regular expressions for handling specific symbols and punctuation;
- to switch from rule-based to neural edit-tree lemmatization;
- to create custom modules for sentence splitting and for word sense disambiguation.

2. Related work

Savkov et al. [19] are the first to present a linguistic processing pipeline for Bulgarian including morphological analysis, lemmatization, and syntactic analysis of Bulgarian texts. Lemmatization and word sense disambiguation are performed by manually crafted rules, while part-of-speech tagging and morphological tagging are performed by tools based on support vector machines (SVMs). Different parts of the pipeline are developed as part of different systems, including the CLaRK system [20], Gtagger [6], and MaltParser [14].

Later, Popov et al. [18], present a spaCy-based pipeline consisting of modules for tokenization, lemmatization, POS tagging, dependency parsing, named entity recognition and dependency parsing. The lemmatization is still done by applying a set of rules over a large morphological dictionary, part of speech tagging, dependency parsing and named entity recognition are done by spaCy's built-in algorithms and word sense disambiguation is performed using the EWISER system [3]. There are some directions in which the pipeline can be improved. First, the pipeline is implemented in an outdated version of spaCy, it is not publicly available, and there is no discussion of the algorithms in use. Second, there is no information about the use of pretrained word vectors. And finally, there are no quantitative evaluation

[1] https://blog.netcetera.com/macedonian-spacy-f3c85484777f

results reported from the WSD system, and there is no analysis of the errors that the system produces.

3. Training data

3.1. Universal dependencies

The data in the Bulgarian treebank [16] consists of a total number of 11138 sentences, of which 8907 are in the train set, 1115 are in the development set, and 1116 are in the test set. These sets are formed in the following way: each first sentence is taken for the test set, each tenth one is taken for the development set, and the rest of the sentences form the train set. The data is from three main domains: 81% from Bulgarian newspapers, 16% from fiction texts, and 3% from administrative documents.

The texts of the dataset are split into sentences, and every sentence is tokenized. Then, every token is annotated with its lemma, part of speech tag, list of morphological features, head of the dependency, type of dependency relation, and other additional characteristics. The data is publicly available[2] in CoNNL-U format.[3]

3.2. BulNet

WordNet [11] is a lexical database organized on the basis of semantic features. Its development began in 1978 [21] for the English language. Currently, there are versions for over 200 languages. It is organized around the idea of synsets – cognitive synonym sets representing sets of words of the same part of speech that can be used interchangeably in a certain context. Each synset has its own short description, often referred to in the literature as a gloss, as well as a list of short examples, illustrating the exact use of the given meaning in a sentence. The Bulgarian version of WordNet is called BulNet [9].

Current BulNet consists of 22092 nouns, 9043 verbs, 8969 adjectives, and 1692 adverbs, which makes it several times smaller in volume than the English version.

When synonymous sets are separated from each other by extremely small and difficult-to-notice marks, the task becomes too complicated, even for humans. There are situations where different annotators would choose different meanings and an agreement of about 80% is achieved, which is perceived as target accuracy for a word sense disambiguation system [3].

3.3. fastText vectors

The pretrained word vectors that we are using are fastText vectors [4] for the Bulgarian language[4]. These vectors are of dimension 300 and are trained on Bulgar-

[2] https://github.com/UniversalDependencies/UD_Bulgarian-BTB
[3] https://universaldependencies.org/format.html
[4] https://fasttext.cc/docs/en/crawl-vectors.html

ian Wikipedia. FastText vectors showed the best performance among other architectures for pretrained vectors that we experimented with. One of the main reasons is that fastText works on a character level, generating n-grams of symbols on which the algorithm is trained. This makes it very well suited for morphologically rich languages, such as the Bulgarian language, where a single word can have many forms, not all of which are present in the training corpus. With the n-grams mechanism, FastText is also able to manage successfully out-of-vocabulary cases, which occur often when the training data is limited.

4. Pipeline implementation

In spaCy, there are two types of components - trainable and non-trainable. Trainable components rely on training data and machine learning algorithms. Non-trainable components rely on predefined sets of rules, by which they process the data. In our pipeline, the rule-based components are the Tokenizer and the Sentencizer. The remaining components rely on training data and machine learning modules. In the following subsections, we will discuss in more detail the algorithms which each one of the components is using. Figure 1 presents the sequence of the steps in the pipeline.

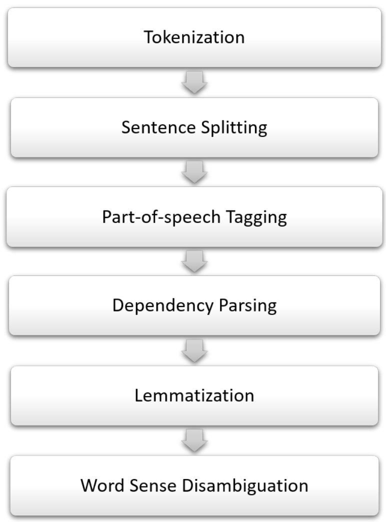

Figure 1. Sequence of the steps of the developed Bulgarian language pipeline

4.1. Rule-based components

4.1.1. Tokenizer

The tokenization is the first step of the pipeline, as the following components need input data in the form of a sequence of tokens. For the sake of modelling correctly the Bulgarian language, we created our own custom tokenizer. The Bulgarian tokenizer consists of:

- Lists of special cases, such as metrics, abbreviations, and titles,
- List of stop words,
- Regular expressions for handling tokens with special symbols, like hyphens, apostrophes,
- Regular expressions for handling punctuation.

Tokenizer exceptions consist of the following types:

- Units of measure: can be with ("см.") or without dot ("см"),
- Abbreviations - some like "изд." (from "издателство", *publishing house*) can be only in the middle of a sentence, whereas others like "др." (from "други", *others*) can be both in the middle and in the end,
- Hyphenated abbreviations, like "г-жа" (*Mrs.*),
- Capitalized abbreviations, like "БДС" (from "Български държавен стандарт", *Bulgarian State Standard*).

4.1.2. Sentence splitter

Ideally, the dependency parser algorithm should be able to learn to split sentences automatically. Unfortunately, as our training data is already split into sentences, that was not possible, and a separate rule-based algorithm had to be developed.

The sentence splitter consists of rules for treating punctuation and a variety of edge cases, connected to the uses of initials and abbreviations. Initials, such as "A.", are marked as invalid end of sentence.

This custom-built module takes as input the tokenized text. In order to split the sentences correctly, the algorithm assumes that a token can be the beginning of a sentence if:

- Starts with an uppercase letter, and
- The preceding token is not an invalid end, and
- The preceding token is end-of-sentence punctuation, or it's not one of the special cases, or is of the special cases, but is a possible end of the sentence.

In this manner, the Sentencizer is able to avoid splitting sentences where the dot is used in abbreviations, such as:

Св. Николай Чудотворец е роден 15 март 270 г. в Патара, Ликия.
(St. Nicholas the Wonderworker was born on March 15, 270 in Patara, Lycia.)

4.2. Built-in trainable components

4.2.1. POS-tagger and morphologizer

The part of speech tagger and morphologizer components are implemented by the spaCy's tagger model, which uses a linear layer with softmax activation to predict tag scores for every token's vector. The POS tagging module uses as features the token vectors, as well as information from the morphologizer, which is a trainable component that predicts morphological features and fine-grained POS tags following the Universal Dependencies UPOS[5] and FEATS[6] annotation guidelines.

4.2.2. Dependency parser

A dependency parser (DEP) is a model which analyzes the grammatical structure of a sentence. The dependency parser marks the relationships between "head" words and words that modify those heads. The spaCy parser uses a modification of the non-monotonic arc-eager transition system [7], which jointly learns dependency parsing and labelled dependency parsing. The algorithm uses a pseudo-projective dependency transformation [13], which allows it to work with non-projective trees, which may occur in languages with free word order, such as the Bulgarian language. In our training data, 279 out of 8836 sentences (3%) have non-projective dependency trees.

4.2.3. Lemmatizer

In the Bulgarian language, the lemma of a certain word cannot be determined by applying a short list of rules. One approach to the problem is to use large lists of words in all their possible forms and base form to determine the result of lemmatization. A disadvantage of the approach is that when the desired word does not appear in the list, there is no processing option. Then either the current form of the input word should be returned or an empty character string should be returned. If the dictionary of presented word forms is not comprehensive enough, both approaches will result in token mishandling and, therefore, in low recall values of the overall system.

To address this limitation, we apply a method proposed by Müller et al. [12], according to which an "EditTreeLemmatizer" is built. The method is available for use as a standalone component in the spaCy tool. In the Neural edit-tree lemmatization algorithm, the lemmatization task is treated as a classification problem. The classes represent all learned edit trees, and the Softmax function is used for computing the probability distribution over all trees for a particular token. Then the algorithm tries to apply the most probable tree and, if this is not possible, continues with the next most probable tree. An edit tree consists of the following types of nodes: inferior nodes, which split the string into a prefix, an infix, and a suffix, and leaf nodes, which apply the learned transformation. An example edit tree is shown in Figure 2.

[5] https://universaldependencies.org/u/pos/index.html
[6] https://universaldependencies.org/format.html#morphological-annotation

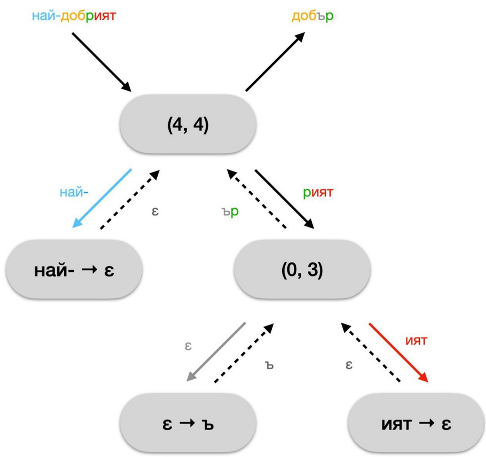

Figure 2. Edit tree for the inflected form “най-добрият“ (*the best*) and its lemma “добър” (*good*)

4.3. Additional components

4.3.1. Word sense disambiguation

Word-sense disambiguation (WSD) is the task of determining the correct sense of a word in a given context [8]. It can be a supervised, unsupervised, or hybrid task. Supervised approaches utilize a corpus of sentences in which individual words are manually labelled with senses from a lexical resource, such as WordNet.

In the current pipeline, word sense disambiguation is regarded as a supervised machine-learning task. Bulgarian WordNet provides a valuable source of data, including possible senses, hypernyms, and usage examples. For each potential sense, a sample text is constructed by combining the glosses for the sense and its related hypernyms. Subsequently, these texts are passed through spaCy as distinct documents. This process results in vectors of each token in the text. Next, the vectors corresponding to each document are averaged, and each sentence is assigned a single vector. The rationale behind this approach is that different sense descriptions will have distinct vector representations, and by averaging them, the representation in the latent space will contain sufficient information for disambiguation. The final stage entails identifying the closest candidate to the target sentence in the latent space, which is accomplished by employing cosine distance for measuring distance and determining the closest candidate.

5. Experiments

5.1. Word embeddings

For the performance of the word sense disambiguation model, which uses information from the preceding steps of the pipeline, we experimented with different pretrained word embeddings architectures, such as BERT [5], RoBERTa [10], Flair [2] and fastText [4]. Bulgarian BERT[7] is trained on the Bulgarian version of Wikipedia and the OSCAR 2019 corpus [15], composed of 1,268,114,977 words with a total size of 14 GB, as well as with data from the online library `Chitanka.info`[8]. Bulgarian RoBERTa[9] is trained on the corpora Wikipedia, OSCAR 2019, and Newscrawl one million sentences[10].

An interesting observation is that for our task BERT and RoBERTa models have similar performance. This is most likely due to the volume of data on which the Bulgarian models were trained. For the original English models, BERT was trained on 16 GB of data and RoBERTa – on 160 GB. Such volumes of data are not available for the Bulgarian language, and accordingly the models have a comparable performance.

In our experiments, the best results were obtained with the fastText vectors, which we included in the final version of the system. A comparison of the results can be found in Table 1.

Table 1. Performance comparison of different approaches for solving the WSD task

Algorithm	Accuracy
Cosine similarity with fastText	65.24
Cosine similarity with Flair	63.99
Cos similarity with RoBERTa	62.82
PageRank – undirected graph	58.76
PageRank – directed graph	61.33

5.2. Word sense disambiguation algorithms

For the word sense disambiguation task, we experimented with two main types of algorithms – graph-based and similarity-based. One of the tested methods for resolving ambiguity involved using graph-based algorithms. This family of algorithms extracts all possible senses for a given text and builds graph representation of the text utilizing the available relations in WordNet or other similar knowledge bases. Those approaches rely heavily on the PageRank algorithm, with different modifications tailored to suit the specific problem at hand. In essence, all approaches assign weights to the nodes in the graph, which determine the most probable senses for the target text. According to Agirre et al. [1], such methods need longer texts of no less than 20 words to build a graph with weights that will ultimately converge to a single probable sense.

A crucial factor influencing the system's performance is the number of connections and senses available in the knowledge base. In the case of Bulgarian language WordNet, the available connections are fewer and lack utility for graph construction,

[7] https://huggingface.co/rmihaylov/bert-base-bg
[8] https://chitanka.info/
[9] https://huggingface.co/iarfmoose/roberta-base-bulgarian
[10] https://wortschatz.uni-leipzig.de/

since the sentence graphs may not be connected and can be composed of several disjoint graphs. As for the similarity-based approaches, the number of relations in WordNet is not significantly important because in those methods we rely heavily on the semantic relationships that have been learned from unstructured texts while training the word embeddings. Building word embeddings does not require the meticulously crafted relations encapsulated in WordNet and the embeddings' position in latent space can be used to calculate the similarity between words and sentences. In our case, the cosine similarity was used to determine the distance between different sentences. In order to deal with different sentence lengths, an average pooling was used that averages the weights of the words within the provided excerpt, resulting in an effective assignment of a single word embedding for each sentence.

Our experiments indicated that vector-based approaches prove more effective for the Bulgarian language. Table 1 summarizes the results of the different word disambiguation techniques tested.

6. Results and evaluation

6.1. Automatic evaluation

We evaluate our pipeline, based on the following metrics: TOK – tokenization accuracy, POS – part-of-speech accuracy, UAS – unlabeled attachment score for dependency parsing, LAS – labelled attachment score, LEMMA – lemmatization accuracy, WSD – word sense disambiguation accuracy, and SENT F – sentence splitting F-score. The sentence splitting score, however, is biased, as the majority of the test examples are already split into sentences.

Table 2 presents the results of the current pipeline, compared to the previously reported results [18]. We observe that our pipeline archives better results on all the included subtasks.

The previous works do not report automatic evaluation for the lemmatization and word disambiguation task, and we could compare the obtained results. However, manual error analysis, presented in the next section, shows that those two modules also show good performance for the Bulgarian language.

Table 2. Comparison of the results on the BulTreeBank dataset of the current pipeline and the previous implementation [18]

Metric	Pipeline-2020	Pipeline-2023
TOK	—	99.97
POS	94.49	98.12
LEMMA	—	93.88
UAS	89.71	89.95
LAS	83.95	84.77
WSD	—	65.24
SENT F	—	94.65

6.2. Error analysis

6.2.1. Lemmatization

A frequently occurring problem are the differences in the spelling of the base forms of the words. In BulTree Bank, the data on which the lemmatizer was trained differs from the way it is written in WordNet. Such examples are "запазя-(се)" – "запазя" (*to save oneself*), "смея се" – "смея" (*to laugh*) and others. During the system workflow, the suffixes "-(се)" and "-се", which are marking reflexive verbs, are removed before the search is executed. Unfortunately, this does not help in cases where the proposed base form is wrong. Such a case is "призовам-(се)" instead of "призовавам" (*to call out*).

Alignment of the resources would lead to an improvement of the results and an increase of the recall measure and, accordingly, of the F1 result of the overall system. On the BulTree Bank validation set an accuracy of 0.9388 is reached for the lemmatization task, which is quite a good result for a morphologically rich language such as Bulgarian. Despite the high score on this set, discrepancies with WordNet undoubtedly contribute to low recall values.

We examined closely the outputs of the edit tree lemmatization algorithm. We were able to identify three main types of mistakes made by the algorithm:

1. Errors caused by the suffixes "-(се)" and "-се".
2. Incorrect suggestion for a base form of an existing word, such as "булеварда" instead of "булевард" and "пейзажа" instead of "пейзаж".
3. Prediction non-existent words, such as "лип" instead of "липа", and "щайг" instead of "щайга".

As we can see, most of the errors come from confusion between words in the masculine gender, ending with the determination ending *-а*, and words in the feminine gender, whose base form ends with a gender ending *-а*. It is possible that with more training examples the algorithm for tree selection will be able to discern more details and choose the appropriate tree more frequently. In its current version, the model is trained on 8,907 sentences from the BulTree Bank, which is insufficient for a morphologically rich language such as Bulgarian.

6.3. Word sense disambiguation

There are two main sources of errors of the word sense disambiguation module:

1. Synsets, which are presented in BulNet with multi-token expressions, such as "черен чай" (*black tea*) and "маслодайни рози" (*oil roses*).
2. Overlapping senses – often there are cases where the meaning of a word in a particular sentence can fall in more than one of the predefined senses, and the predicted and original senses are different, while equally true. An example of such a case is shown in Table 3.

Table 3. Example of an error by the word sense disambiguation module for the meaning of the word "пристъпвам" (*to step*), where the predicted and the original sense are very close

Example	Expected sense	Predicted sense
Пристъпих напред и вдигнах ръка. (I stepped forward and raised my hand.)	*btbwn-038000141-v* Движа се като правя стъпка след стъпка в равномерен ритъм. (I move by taking step after step in a steady rhythm.)	*btbwn-038000146-v* Правя, направям една или няколко стъпки, обикновено в посоката, към която гледам, към която съм обърнат. (I take one or more steps, usually in the direction I'm looking or, facing.)

7. Conclusion

We presented the implementation of an open-source pipeline for processing the Bulgarian language, built on the natural language processing library spaCy, which shows significant results on numerous tasks. We systemized the available lists of tokenizer exceptions and successfully created new custom modules for sentence splitting and word sense disambiguation and a new neural-based module for lemmatization. Finally, we released an open-source version of the pipeline.

The presented pipeline can be used in multiple ways, some of which include: in sentiment analysis and hate speech detection tasks, by lemmatizing text and searching in a list of predefined signaling words; in machine translation, to find the right meaning of an ambiguous word and produce the right translation; in text categorization tasks, by providing additional information about the text, such as additional features of the words and sentences of its contents. These and any other applications can be built by appending the pipeline with additional components (such as one for text categorization) or integrating it with other systems.

7.1. Limitations

There are several ways in which the work on the pipeline can be improved.

First, the sentencizer can be further developed to process nested sentences, typical for the press and literature. For this, additional linguistic knowledge will be needed in order to model more complicated cases.

The lemmatization module can benefit from additional data processing, and also from more data, as a significant number of edit trees modelling the lemmatisation process of Bulgarian words can exist.

Currently, our method for word sense disambiguation is unable to process word bigrams and trigrams. When the target phrase consists of more than one word, the overall model fails because the search is only performed on a single token. To improve the word sense disambiguation module performance, it is necessary to implement a comprehensive system to deal with bigrams and trigrams. Their presence in WordNet makes searching difficult in situations where there is no complete match and searching by lemma form does not lead to successful detection. Additionally, further preprocessing is needed to solve the problem of the difference in the word forms used in the two datasets – BulTreeBank and WordNet.

7.2. Ethics Statement

This work aims at an equal and fair distribution of language technologies for different populations, especially speakers of low-resource languages, such as Bulgarian. The pipeline will be publicly available for use and modification.

We consider that there are no potential harms for different groups, if the described tool is misused. On the contrary, such a technology can help in analyzing and filtering propaganda, misinformation and hate speech in texts.

Acknowledgements

An initial version of the sentencizer and the lists of the tokenizer exceptions were compiled by Luboslav Krastev and Daniel Traikov from Apis Europe, with the assistance of Prof. Dr. Kiril Simov from the Institute of Information and Communication Technologies of the Bulgarian Academy of Sciences. These resources, as well as the data from the Bulgarian WordNet, are provided with the kind assistance of Prof. Dr. Kiril Simov.

References

[1] E. Agirre, O. L. de Lacalle and A. Soroa, Random walks for knowledge-based word sense disambiguation, Comput. Linguist. 40(1) (2014) 57–84, https://aclanthology.org/J14-1003.

[2] A. Akbik, T. Bergmann, D. Blythe, K. Rasul, S. Schweter and R. Vollgraf, FLAIR: An easy-to-use framework for state-of-the-art NLP, in: Proc. 2019 Conf. of the North American Chapter of the Association for Computational Linguistics (Demonstrations), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, 54–59, https://aclanthology.org/N19-4010.

[3] M. Bevilacqua and R. Navigli, Breaking through the 80% glass ceiling: Raising the state of the art in word sense disambiguation by incorporating knowledge graph information, in: Proc. 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, 2854–2864, https://aclanthology.org/2020.acl-main.255.

[4] P. Bojanowski, E. Grave, A. Joulin and T. Mikolov, Enriching word vectors with subword information, Trans. Assoc. Comput. Linguist. 5 (2017) 135–146, https://arxiv.org/abs/1607.04606.

[5] J. Devlin, M.-W. Chang, K. Lee and K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: Proc. 2019 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, vol. 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, 4171–4186, https://aclanthology.org/N19-1423.

[6] G. Georgiev, V. Zhikov, A. Ontotext, P. Osenova, K. Simov and P. Nakov, Feature-rich part-of-speech tagging for morphologically complex languages: Application to Bulgarian, EACL 2012 (2012) 492, https://arxiv.org/abs/1911.11503.

[7] M. Honnibal and M. Johnson, An improved non-monotonic transition system for dependency parsing, in: Proc. 2015 Conf. on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Lisbon, Portugal, 2015, 1373–1378, https://aclanthology.org/D15-1162.

[8] D. Jurafsky and J. Martin, Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition, Prentice Hall Ser. in Artificial Intelligence, Pearson Prentice Hall, 2009, https://books.google.bg/books?id=fZmj5UNK8AQC.

[9] S. Koeva, S. Mihov and T. Tinchev, Bulgarian wordnet – structure and validation, Romanian J. Inf. Sci. Technol. 7(1–2) (2004) 61–78, http://lml.bas.bg/~stoyan/wordnet_logic.pdf.

[10] Y. Liu, M. Ott, N. Goyal et al., Roberta: A robustly optimized bert pretraining approach, arXiv preprint arXiv:1907.11692 (2019), https://arxiv.org/abs/1907.11692.

[11] G. A. Miller, Wordnet: A lexical database for English, Commun. ACM 38 (1995) 39–41, https://doi.org/10.1145/219717.219748.

[12] T. Müller, R. Cotterell, A. Fraser and H. Schütze, Joint lemmatization and morphological tagging with lemming, in: Proc. 2015 Conf. on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Lisbon, Portugal, 2015, 2268–2274, https://aclanthology.org/D15-1272.

[13] J. Nivre and J. Nilsson, Pseudo-projective dependency parsing, in: Proc. 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05), Association for Computational Linguistics, Ann Arbor, Michigan, 2005, 99–106, https://aclanthology.org/P05-1013.

[14] J. Nivre, J. Hall, J. Nilsson, A. Chanev, G. Eryigit, S. Kübler, S. Marinov and E. Marsi, Maltparser: A language-independent system for data-driven dependency parsing, Nat. Lang. Eng. 13 (2007) 95–135, http://lrec-conf.org/proceedings/lrec2006/pdf/162_pdf.pdf.

[15] P. J. Ortiz Suárez, L. Romary and B. Sagot, A monolingual approach to contextualized word embeddings for mid-resource languages, in: Proc. 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, 1703–1714, https://aclanthology.org/2020.acl-main.156.

[16] P. Osenova and K. Simov, Universalizing BulTreeBank: a linguistic tale about glocalization, in: The 5th Workshop on Balto-Slavic Natural Language Processing, INCOMA Ltd. Shoumen, Bulgaria, Hissar, Bulgaria, 2015, 81–89, https://aclanthology.org/W15-5313.

[17] E. Partalidou, E. Spyromitros-Xioufis, S. Doropoulos, S. Vologiannidis and K. Diamantaras, Design and implementation of an open source Greek POS tagger and Entity Recognizer using spaCy, WI '19: IEEE/WIC/ACM Int. Conf. on Web Intelligence, 2019, 337–341, https://doi.org/10.1145/3350546.3352543.

[18] A. Popov, P. Osenova and K. Simov, Implementing an end-to-end treebank-informed pipeline for Bulgarian, in: Proc. 19th Int. Workshop on Treebanks and Linguistic Theories, Association for Computational Linguistics, Düsseldorf, Germany, 2020, 162–167, https://aclanthology.org/2020.tlt-1.14.

[19] A. Savkov, L. Laskova, S. Kancheva, P. Osenova and K. Simov, Linguistic processing pipeline for bulgarian, in: Proc. 8ht Int. Conf. on Language Resources and Evaluation (LREC'12), 2012, 2959–2964, http://www.lrec-conf.org/proceedings/lrec2012/pdf/829_Paper.pdf.

[20] K. Simov, Z. Peev, M. Kouylekov, A. Simov, M. Dimitrov and A. Kiryakov, Clark-an xml-based system for corpora development, in: Proc. of the Corpus Linguistics

2001 Conference, 2001, 558–560, http://bultreebank.org/wp-content/uploads/2017/04/BTB-TR06.pdf.

[21] P. Vossen, Wordnet, eurowordnet and global wordnet, Revue française de linguistique appliquée 7 (2002) 27–38, https://doi.org/10.3917/rfla.071.0027.

Received on March 30, 2023
Accepted on May 7, 2023

Melania Berbatova and Filip Ivanov

Faculty of Mathematics and Informatics
Sofia University "St. Kliment Ohridski"
5 James Bourchier Blvd.
1164 Sofia
BULGARIA

E-mails: msberbatov@fmi.uni-sofia.bg
phil.ivanov@outlook.com

ГОДИШНИК НА СОФИЙСКИЯ УНИВЕРСИТЕТ „СВ. КЛИМЕНТ ОХРИДСКИ“
ФАКУЛТЕТ ПО МАТЕМАТИКА И ИНФОРМАТИКА
Том 110

ANNUAL OF SOFIA UNIVERSITY “ST. KLIMENT OHRIDSKI”
FACULTY OF MATHEMATICS AND INFORMATICS
Volume 110

INCREASE DIGITAL HEALTH LITERACY WITH THE PRACTICAL IMPLEMENTATION OF AN E-LEARNING SYSTEM ACCORDING TO WCAG COMPLIANCE STANDARDS

NATASHA BLAZESKA-TABAKOVSKA, SNEZANA SAVOSKA, ANDRIJANA BOCEVSKA, ILIJA JOLEVSKI, BLAGOJ RISTEVSKI and VLADIMIR TRAJKOVIK

Individuals with cognitive difficulties (CD) have very limited access to appropriate learning resources. Their unique individual needs and requirements prevent them from accessing traditional online formal and informal learning methods, and resources are usually inadequate. We develop an e-learning platform with health content as an opportunity for all to increase digital health literacy, including people with cognitive disabilities. The paper aims to present some feature's settings important for people with a wide range of cognitive disabilities.

Keywords: cognitive disability, accessibility, Learning Management System (LMS), ATutor, usability

CCS Concepts:

• Applied Computing∼Education∼E-learning

1. Introduction

One of the general principles that underlie the Convention on the Rights of Persons with Disabilities is *providing access to all aspects for all people on an equal basis.* It means “universal design”, design usable *by all people to the greatest extent possible without the need for adaptation or specialized design.*

Online educational platforms or e-Learning platforms, also defined as Web Based Training, enable users to learn whenever they want, wherever they want, and

DOI: 10.60063/GSU.FMI.110.51-62

what they want – a wide range of topics including health topics. However, accessibility and usability are not often considered when designing e-Learning Platforms or other digital interfaces [27]. Taking into consideration the trend of an increasing rate of people with cognitive disabilities, it is necessary to have a more proactive role in the community to facilitate the accessibility of digital assets and web content.

Cognitive disabilities are term that refers to a broad range of limitations or challenges in performing one or more types of cerebral tasks. For instance, individuals with a cognitive disability may experience difficulty in understanding or processing information, solving problems or responding to stimulus. Center for Persons with Disabilities, Utah State Univ., WebAIM – Web Accessibility in Mind [19], suggests that cognitive disability is defined as having "greater difficulty with one or more types of mental tasks than the average person".

Cognitive disabilities include intellectual disabilities (significant limitations in both intellectual functioning and adaptive behaviour), pervasive developmental disabilities (delays in the development of socialization and communication skills), neurodegenerative disease (all types of traumatic brain injuries caused by external forces, and also brain injuries caused after birth by cerebral vascular accidents, and loss of oxygen to the brain), neurodegenerative disease (includes Alzheimer's disease, amyotrophic lateral sclerosis (ALS), Huntington's disease, and Parkinson's disease) and learning disabilities (neurological conditions that interfere with an individual's ability to store, process, or produce information) [14]. Cognitive disability can result in many different types of functional disabilities, so there can be no one-size-fits-all solution to ensuring access to learning material for all.

According to one of the aims of the Cross4all Project [2, 24], increasing health literacy, we develop digital health content "for all", including those with cognitive impairment [2–4, 26]. The paper highlights some important features for people with cognitive impairments and shows real implementation of the contents. The paper is organized as follows. Section 2 depicts cognitive disabilities, the impact of IT applications on the accessibility of information and knowledge for people with cognitive impartments, and the impact of web accessibility and how Web Content Accessibility Guidelines (WCAG) are suitable for people with cognitive impartment. Section 3 describes the e-learning system for increasing health literacy, which we developed within the project Cross4all, focusing on accessible interfaces or environments. Finally, the last section provides concluding remarks.

2. ICT access for people with cognitive disabilities

Accessible ICT – including accessible off-the-shelf ICT can assist a person by providing daily reminders, navigation assistance, and a means of staying in touch with support networks. Accessible ICT is very important for people with cognitive disabilities in education.

The repetition that ICT enables allows learners to engage for more extended periods. Moreover, mainstream devices, such as tablets, have successfully improved not only students' knowledge and performance skills but also their motivation, communication, and social skills [4].

For years, technology for people with disabilities, including those with cognitive disabilities, centered on specialized assistive technology, was frequently expensive and hard to find. However, the most promising trend in today's ICT ecosystem is the movement toward equipment and software that is customizable to the specific needs and abilities of each user. DAC (Disability Advisory Committee of Federal Communication Commission (FCC)) Best Practices recommend that everywhere in the WWW, where it is appropriate, features and functions that allow personalization and customization that facilitate the accessibility and usability of ICT and applications for people with cognitive disabilities should be incorporated. According to the DAC, these measures would promote the principles of universal design and minimize the need for costly and difficult-to-find accessories [11]. The EU has taken measures to promote equal rights and opportunities for persons with disabilities through various directives, regulations, and initiatives. The European Disability Strategy 2010–2020 was one such initiative that aimed to promote inclusion and combat discrimination. It included measures to improve accessibility, employment opportunities, education, and social inclusion for persons with disabilities.

In addition, the authors proposed a classification of technologies for cognition based on the International Classification of Functioning (ICF), allowing a more straightforward identification of what technology could be helpful for a specific cognitive function impairment [17]. Moreover, this will help solutions to be much more generalizable across the heterogeneity of cognitive impairments. More detail can be found at [11]. However, the applicability of specific recommendations depends on any applicable laws or policies in EU countries.

On the other hand, in the domain of education, global and specific cognitive impairments/learning limitations have negative outcomes on the school performance and the achieved level of professional qualification [8]. Hence, this creates the necessity to increase accessible learning environments for people with cognitive impairments. Therefore, e-learning is expected to be one of the critical tools for improving access to education and aiding social inclusion for people with cognitive disabilities.

According to the systematic literature review in [9] there is: a lack of e-learning studies addressing the issue of accessibility for people with cognitive impairments. These studies are mainly focused on design guidelines rather than effectiveness assessment. They identify five families of accessibility function: adaptive systems, game elements, accessible content, virtual agents and accessible interfaces or environments.

In the distant 1997-year Tim Berners-Lee emphasizes the importance of removing the World Wide Web (WWW) accessibility barriers for all people with disabilities. Since then, more efforts have been made to improve Web accessibility. These efforts have generally consisted of three approaches: 1) developing screen readers, 2) specialized Web browsers, and 3) Web design guidelines. Despite these efforts, there is still much to do in this field [9].

WCAG 1.0 were officially published by the World Wide Web Consortium, Web Accessibility Initiative, in 1999 to encourage Web accessibility [30]. In 2009 W3C recommended new and updated content, WCAG 2.0, in 2018 new and updated content WCAG 2.1 and most recently W3C Candidate Recommendation Draft WCAG 2.2 (01.2023) [29]. Following these guidelines will make content accessible to a broader range of people with disabilities and make Web content more usable to users in general. WCAG 2.0 as standard is classified as the ISO/IEC 40500 standard and contains three implementation levels, 'A', 'AA' and 'AAA'.

According to the Federal Communications Commission, the specific functional needs of many people with cognitive disabilities are: functional difficulties in memory can impact the ability to recall what a person has learned over time; organizational deficits to perform tasks at the appropriate times; various ways of processing information; functional difficulties in problem-solving; difficulty focusing attention; a broad spectrum of difficulties in understanding a text, ranging from mild reading challenges to illiteracy; difficulties in describing their physical locations [14]. According to their extensive literature review, the authors of [16] point out that during the past several years, many steps have been taken to address Web accessibility for people with cognitive disabilities. However, despite this progress, Web accessibility for users with cognitive disabilities lags far behind the general population and behind Web access for other disability groups. Some of the reasons for this lack of progress are the variety of needs specific to the types of disabilities encompassed by the definition of cognitive disabilities. A list of top web access design recommendations for users with cognitive disabilities based on frequency cited by existing web design guidelines are given. The top 22 ranked design recommendations covered text size and shape, consistency of navigation and page design, icons, pictures, text writing, style, margins, hyperlinks, line spacing, and screen layout.

2.1. Learning management system for cognitive disability

The learning management systems, such as ones compatible with life-long training and ones allowing people to gain new skills and knowledge, provide excellent opportunities for creating learning environments for people with disabilities, notably those with cognitive impairments and limited learning activities. These systems offer flexibility so users can adapt their training program to meet their needs. Choosing the most suitable Learning Management System (LMS) for people with a cognitive disability is very important. Over the past few years, commercial LMS vendors, and open-source communities have invested significant resources into making their products more accessible. At the same time, accessibility remains a challenge for users with disabilities and older people [3].

In [25] areas for design considerations which should be taken into account when creating web content that can be used by users with cognitive impairment, are emphasized. They include:

1. The navigation is standardized and unambiguous; Headings, subheadings, and lists for structuring the information are used; menus are short and easy to understand;

2. Provide information in multiple formats, with emphasis on visual format; Supplemental media such as illustrations, icons, and video;

3. Short, simple, unambiguous phrases are easier to understand than long, complex, ambiguous ones;

4. The layout, which covers the availability of screen elements, form or announcement, needs to be consistent on pages, fonts, colors, and locations of page elements; high contrast between text and background needs to be present.

Proactively seeking out and maintaining collaborative relationships with people with cognitive disabilities (either individually or through organizations) that have established expertise with or represent these individuals are essential in all phases: research, design and development of communications products and services [21].

Authors of [23] suggest that courses should be designed to be accessible from the beginning. Implementing universal design principles at the outset avoids costs caused by the need to engage in a digital retrofit and serves to include those students who would otherwise be excluded by an unwillingness to request accommodations. Following [12], access for users with a disability needs to be built into the design process at the beginning, not retrofitted.

Technologies such as web-based lecture systems are valuable for students with disabilities and the broader student population [13]. The text is made available as an audio file and can be listened to in different settings. Subtitles can be used to read the content of a video presentation when sound is not appropriate. Information that is less fixed to a specific format can be accessed in multiple ways and is more easily searchable [20].

Despite the various possibilities offered by LMS to overcome the functional difficulties of people with cognitive disabilities and despite adopting a compliance approach WCAG 1.0 and WCAG 2.0 from the W3C, there is still a tendency to see accessibility as an afterthought or a potential legal liability to overcome. The authors of [10] critiqued learning management systems for adopting a one-size-fits-all approach to accessible e-Learning through adopting a compliance approach WCAG 1.0 and WCAG 2.0 from the World Wide Web Consortium (W3C). Regarding web accessibility, the work produced by the W3C Cognitive and Learning Disabilities Accessibility Task Force is an interesting approach for improving accessibility for people with cognitive impairments.

Part of the problem related to accessibility through LMS is that some LMS allow each person designing a course to set out their own web interface. For example, the user can overwrite inconsequential images with alternative text that clutters the page unnecessarily. Therefore, on the page, there are many unnecessary elements that make navigation difficult. In addition, the learning methods implemented by e-Learning platforms are not always practical for people with functional cognitive disabilities. People with disabilities may require different ways to interact with digital content in courses on e-Learning platforms. In [15] it is pointed out that even with an accessible basis for the LMS, the course content hosted through these

LMS may have accessibility issues. According to [18], "disabled students can access the e-learning platform but not contents, resources, activities, collaboration and interaction tools".

In [7] it is emphasized that the basic accessible framework is underpinned by the seven principles of Universal Design (UD), the three principles of Universal Design for Learning (UDL), and the four principles of the Web Content Accessibility Guidelines (WCAG). UD is defined by the Center for Universal Design as "the design of products and environments to be usable by all people, to the greatest extent possible, without the need for adaptation or specialized design" [7]. Principles for the UD of any product or environment include the following: equitable use, flexibility in use, simple and intuitive use, perceptible information, tolerance for error, low physical effort, and appropriate size and space for approach and use.

To overcome the limitations, an ability-based design framework is presented in [28]. The framework is built on three principles:

1. STANCE: Designers should focus on what a person can do instead of what a person cannot do and address accessibility issues by designing flexible systems that can be adapted to people without having to alter their bodies, knowledge, or behavior.

2. INTERFACE: User interfaces may be self-adaptive or user-adaptable to ensure the best possible match between user's abilities and the required use.

3. SYSTEM: Through the monitoring of user actions, the system may model and predict changes in user performance depending on the context and adjust their parameters to compensate [28].

3. Cross4all e-learning system for increasing health literacy

In order to meet IPA2 (Interreg IPA CBC EU Programme) – Cross4all project aim to increase e-health and digital health literacy [24], we used e-learning platform with accessibility and all previously stated conclusions in mind. The application of the WCAG provided by the W3C in the improvement of health and social services is crucial for the broader population as it can significantly contribute to increasing the level of digital e-health literacy [3].

Although there are many easy-to-use, user-friendly developed LMS algorithms to help make choosing the best LMS easy, choosing the appropriate LMS in accordance with the WCAG standard, that aimed to increase the digital and health literacy of the population of the cross-border area in Cross4all project of IPA2, was a challenging task. We have done more research on relevant literature and practical research and analysis to answer this challenge. We studied and used known standards in healthcare [4, 22] to use a practical LMS for this purpose. More details for this analysis can be found in [3, 4].

Firstly, we analyze five Learning Management Systems (Moodle, Eliademy, Docebo, Sakai and Atutor) that are most used by the community according to Gartner, which does not mean that they are the best for all e-learning areas. We aimed to analyse the accessibility of LMSs for people with disabilities considering the criteria with different levels of compliance in accordance with the WCAG 2.0 standard [5]. We realize that each of the analysed LMSs is with unique features. Each LMS has core features that cannot be modified and variable features which can be adapted for the specific accessibility requirements of people with disabilities. According to our conclusions, all of the evaluated LMSs except Docebo satisfy the criteria according to WCAG 2.0 standards for Level A. Regarding WCAG 2.0 standards Level AA only Moodle and ATutor fulfil the established criteria. This analysis was expanded in accordance with the additional demands of WCAG 2.1 standard, which includes 17 new success criteria: five for Level A, seven for Level AA and five for Level AAA of compliance. Most of the criteria are fulfilled [6]. The comparison tables can be seen at [5, 6] as well as the explanation why the ATutor platform is chosen.

We chose ATutor as an e-learning platform for health topics. Besides the accessibility features, which are the most important attributes for people with disability and older people, it has the best communication tools with a user-friendly interface and encapsulates multimedia help of the Handbook. Moreover, ATutor has the user-friendly JavaScript WYSIWYG editor, which is appropriate for a wide community of users and many valuable features different from others, such as a glossary, users online and search bar, latest discussion topics on forum and sitemap with an ARIA tree [3].

The implemented e-learning platform[1] contained much public content: manuals about the new workflow and mobile applications, tutorials, technical support, and health content. All contents are available in three languages: English, Macedonian and Greek. The users can register themselves, log in on Cross4all LMS and improve their e-health literacy in many areas such as cardiology, psychology, diabetes, how to use gadgets for measuring vital signs, and so on. The selected content can be viewed online or exported/downloaded and viewed offline.

While setting up ATutor and creating web content, which will be used by the users with cognitive impairments, we consider different guidelines and the proposed four design main areas: Navigation, Functionality, Content and Text and Page layout [4].

ATutor has features to add skip-to-content shortcuts so learners can easily access the main content area without repetitive navigation actions when using a keyboard. For example, we considered this when we designed health courses. Also, we made navigation standardized and unambiguous; we used headings, subheadings, and lists for structuring the information; we used short menus that were easy to understand with limited options to prevent cognitive overload (Figure 1).

We provide information in multiple formats. An accessible e-learning course should have different formats of content. Learners with cognitive disabilities might find focusing on long-form written content hard and tiring. These learners can benefit instead from video content. ATutor allows recording or uploading videos easily (Figure 2). The uploaded videos also contain subtitles or transcripts to ensure

[1] http://atutor.cross4all.uklo.edu.mk/login.php

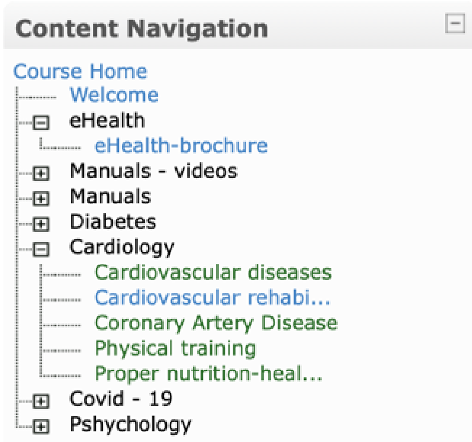

Figure 1. Standardized and unambiguous navigation in Cross4all e-learning system

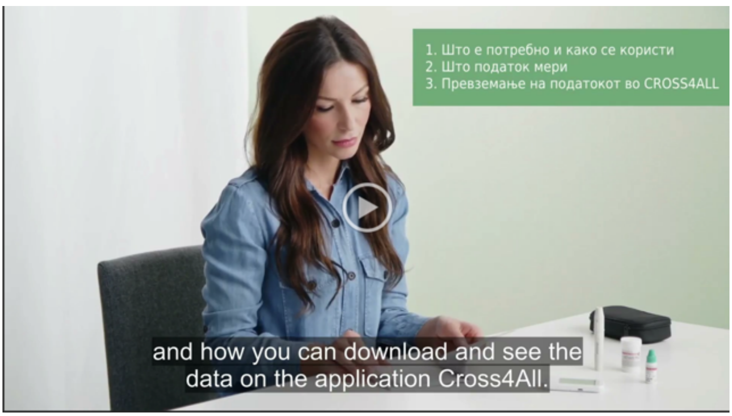

Figure 2. Video contents in Cross4all e-learning system

additional opportunities. Furthermore, we add icons and audio files, which have the potential to significantly enhance the accessibility of web content for people with cognitive disabilities (Figure 3).

Audio content is also an essential element of accessible e-learning courses. ATutor allows uploading audio files easily. This way, learners who cannot follow written content can listen to health-related content.

Screen readers are important when designing accessible e-learning courses. These features are incorporated into ATutor. If a learner with a cognitive impairment needs additional help following the learning content, the screen reader can be activated.

Different fonts and colours make content easy to read and follow for all learners. Zoom-in functions supported in ATutor allow learners to increase the texts and images until they clearly understand the content. Moreover, the text is split into

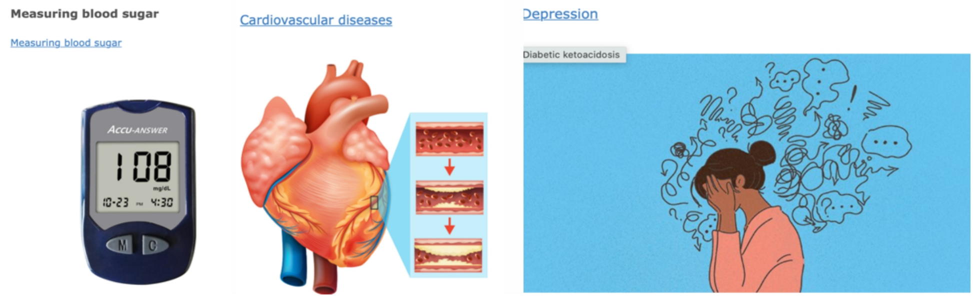

Figure 3. Icons and audio make easier accessibility content's in Cross4all e-learning system

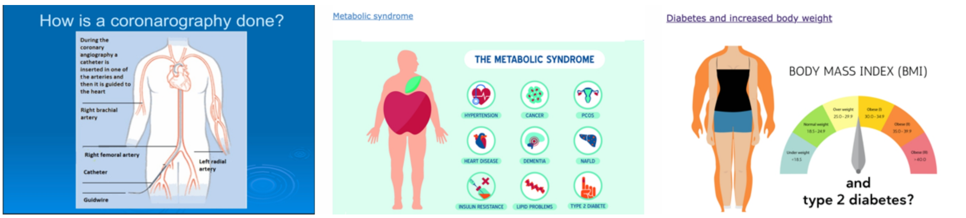

Figure 4. Cross4all e-learning system: contrast, readability, distinguish the words from the background

paragraphs to avoid large chunks of text and to make the text more readable. Finally, fonts that make the text easy to understand and background options that do not make it hard to distinguish the letters are chosen.

We care about contrast; it is not only important for aesthetic reasons. It is also key to making content more readable. Therefore, we choose high-contrast themes, so the learner can easily distinguish the words from the background (Figure 4). Also, we take care of streamlining page design.

Additionally, the users can change their profile elements, enrol in courses they are interested in, and control which versions of content are displayed if, for example, the primary version is not accessible to them, or they prefer an alternate format. Also, users can develop a network of contacts, create and participate in social groups, develop a social profile and use different gadgets (a user can add some of the available applications to their Social Networking environment that provide a whole range of potential networking functionality) and they can create "My Contacts" – a list of people in their social network [1].

4. Conclusion

E-learning is a growing area, and this teaching practice holds great potential to be an avenue of inclusion for people with cognitive disabilities. Although different LMS have different approaches to accessibility, how different platforms and networks are structured will influence and how they can be adapted for greater accessibility.

With the consciousness that an e-learning platform should be accessible to all, all our efforts were put into it. We used an LMS system to increase e-health digital literacy using the WCAG-compliant e-learning platform ATutor for the aims of Cross4all project. We presented some capabilities for people with cognitive disabilities to use the ATutor e-learning platform [1, 3–5], considering some urgent conditions from internal medicine, COVID-19 condition, and psychological conditions. Many other possibilities can be developed in the future, creating some content that provides cognitive help with other methods and techniques.

With opportunities different fonts, zoom-in functions, background options supported in ATutor and with use high-contrast themes and streamlining page design we made content easy to read and follow for all learners. Additionally, we made the text more readable with split the text into paragraphs and avoid large chunks of text.

We provide information in multiple formats. The learners can benefit from video content, subtitles or transcripts, icons, and audio files. All this opportunity has the potential to significantly enhance the accessibility of web content for people with cognitive disabilities. The course content is logically organized and presented in a structured manner. We use clear headings, subheadings, and sections to divide the material. This allows students to easily locate and access specific health topics within the course. Consistency in navigation elements throughout the course is present. The visual cues, such as bolding, colour, or icons, we have used to draw attention to important information or actions. The highlighting key points or navigation options we have used to make them more noticeable and easier to locate. The connection of this system with other digital libraries is also just one of the opportunities to enhance our e-learning platform for increasing digital health literacy, especially for people with cognitive diseases.

References

[1] N. Blazeska-Tabakovska, I. Jolevski, B. Ristevski, S. Savoska and A. Bocevska, Implementation of e-learning platform for increasing digital health literacy as a condition for integration of e-health services with PHR, in: Proc. Information Systems & Grid Technologies (ISGT'2022), CEUR Workshop Proceedings, 335–345, https://ceur-ws.org/Vol-3191/paper29.pdf.

[2] N. Blazheska-Tabakovska, A. Bocevska, I. Jolevski et al., Implementation of cloud-based personal health record integrated with IoMT, in: Proc. Information Systems & Grid Technologies (ISGT'2021), CEUR Workshop Proceedings, 178–188, https://ceur-ws.org/Vol-2933/paper18.pdf.

[3] N. Blazheska-Tabakovska, B. Ristevski, S. Savoska and A. Bocevska, Learning management systems as platforms for increasing the digital and health literacy, in: 3rd Int. Conf. on E-Education, E-Business and E-Technology (ICEBT 2019), August 02–04, 2019, Madrid, Spain.

[4] N. Blazheska-Tabakovska, S. Savoska, B. Ristevski, I. Jolevski and D. Gruevski, Web content accessibility for people with cognitive disabilities, in: 9th Int. Conf. on Applied Internet and Information Technologies AIIT 2019, October 3–4, 2019, Zrenjanin, Republic of Serbia.

[5] A. Bocevska, S. Savoska, B. Ristevski and N. Blazheska-Tabakovska, Analysis of accessibility of the e-learning platforms according to the WCAG 2.0 standard compliance, in: Int. Conf. on Applied Internet and Information Technologies, November 5 2018, Bitola, Republic of North Macedonia.

[6] A. Bocevska, S. Savoska, B. Ristevski, N. Blazheska-Tabakovska and I. Nedelkovski, A comparison of accessible e-learning projects for improving of digital health literacy, in: Proc. Information Systems & Grid Technologies (ISGT'2018), CEUR Workshop Proceedings, 50–60, https://ceur-ws.org/Vol-2464/paper5.pdf.

[7] S. Burgstahler, A tutorial for making online learning accessible to students with disabilities, University of Washington, https://www.washington.edu/doit/tutorial-making-online-learning-accessible-students-disabilities [Accessed March 20, 2023].

[8] P.-A. Cinquin, P. Guitton and H. Sauzéon, Designing accessible MOOCs to expand educational opportunities for persons with cognitive impairments, Behav. Inf. Technol. 40(11) (2021) 1101–1119, https://doi.org/10.1080/0144929X.2020.1742381.

[9] F. E. Chadli, D. Gretete and A. Moumen, A systematic Literature Review, SHS Web of Conferences 119 (2021) 06005, https://doi.org/10.1051/shsconf/202111906005.

[10] M. Cooper and A. Heath, Access for all to eLearning, in: Research Reflections and Innovations in Integrating ICT in Education, ed. by A. Méndez-Vilas, A. Solano Martín, J. A. Mesa González, and J. Mesa González, FORMATEX, Badajoz, Spain, 2009, 1139–1143.

[11] Disability Advisory Committee, https://www.fcc.gov/disability-advisory-committee [Accessed March 20, 2023].

[12] K. Ellis and M. Kent, Disability and New Media, Routledge, 2013, 184 pp, ISBN 9780415835923.

[13] M. Fardon and J. Williams, On-Demand Internet-transmitted lecture recordings: attempting to enhance and support the lecture experience, in: Exploring the frontiers of e-learning: borders, outposts and migration – Proceedings of ALT-C, ed. by J. Cook & D. Whitelock, University of Manchester ed., Vol. 1, Hobbs the Printers Ltd., 2005, 153–161.

[14] Federal Communications Commission, Individuals with Cognitive Disabilities: Barriers to and Solutions for Accessible Information and Communication Technologies, White Paper, Washington, 2016.

[15] C. S. Fichten, V. Ferraro, J. V. Asuncion, C. Chwojka et al., Disabilities and e-learning problems and solutions: an exploratory study, Educational Technology & Society 12(4) (2009) 241–256.

[16] M. G. Friedman and D. N. Bryen, Web accessibility design recommendations for people with cognitive disabilities, Technol. Disabil. 19 (2007) 205–212.

[17] A. Gillespie, C. Best and B. O'Neill, Cognitive function and assistive technology for cognition: A systematic review, J. Int. Neuropsychol. Soc. 18(1) (2012) 1–19, https://doi.org/10.1017/S1355617711001548.

[18] E. Guglielman, E-learning and disability: Accessibility as a contributor to inclusion, in: 5th Doctoral Consortium at the European Conference on Technology Enhanced Learning, Barcelona, Spain, 29 September, 2010, 31–36.

[19] Institute for Disability Research, Policy, and Practice, Cognitive Disabilities, http://webaim.org/articles/cognitive/design [Accessed March, 2023].

[20] M. Kent, Disability and eLearning: Opportunities and barriers, Disability Studies Quarterly (DSQ) 35(1) (2015) https://doi.org/10.18061/dsq.v35i1.3815.

[21] H. Mariger, Cognitive disabilities and the Web: Where accessibility and usability

meet?, The National Center on Disability and Access to Education, http://ncdae.org/resources/articles/cognitive/.

[22] B. Ristevski and S. Savoska, Healthcare and medical Big Data analytics, in: Applications of Big Data in Healthcare, ed. by A. Khanna, D. Gupta and N. Dey, Academic Press, 2021, 85–112, https://doi.org/10.1016/B978-0-12-820203-6.00005-9.

[23] J. B. Roberts, L. A. Crittenden and J. C. Crittenden, Students with disabilities and online learning: A cross-institutional study of perceived satisfaction with accessibility compliance and services, The Internet and Higher Education 14(4) (2011) 242–250, https://doi.org/10.1016/j.iheduc.2011.05.004.

[24] S. Savoska, V. Kilintzis, B. Jakimovski et al., Cloud based personal health records data exchange in the age of IoT: the Cross4all project, ICT Innovations 2020, in: Proc. Machine Learning and Applications: 12th Int. Conf., September 24–26, 2020, Skopje, North Macedonia Springer Int. Publ., 28–41.

[25] Siteimprove Cognitive Disabilities and Web Accessibility, https://www.siteimprove.com/glossary/digital-accessibility/ [Accessed March 20, 2023].

[26] K. Veljanovska, N. Blazheska-Tabakovska, B. Ristevski and S. Savoska, User interface for e-learning platform for users with disability, in: Proc. Information Systems & Grid Technologies (ISGT2020), CEUR Workshop Proceedings, 68–81, https://ceur-ws.org/Vol-2656/paper7.pdf.

[27] V. Voykinska, S. Azenkot, S. Wu and G. Leshed, How blind people interact with visual content on social networking services, in: Proc. 19th ACM Conf. on Computer-Supported Cooperative Work & Social Computing, ACM, 2016, 1584–1595.

[28] J. O. Wobbrock, S. K. Kane, K. Z. Gajos, S. Harada and J. Froehlich, Ability-based design: Concept, principles and examples, ACM Trans. Accessible Computing 3(3) (2010) 9.

[29] World Wide Web Consortium, Web Accessibility Initiative, Web Content Accessibility Guidelines (WCAG) 2.0 (2008) https://www.w3.org/TR/WCAG20/ [Accessed March 20, 2023].

[30] World Wide Web Consortium, Web Accessibility Initiative, Web Content Accessibility Guidelines 1.0 (1999) https://www.w3.org/TR/WAI-WEBCONTENT/ [Accessed March 20, 2023].

Received on April 1, 2023
Accepted on May 7, 2023

Natasha Blazeska-Tabakovska, Snezana Savoska, Andrijana Bocevska, Ilija Jolevski and Blagoj Ristevski

University "St. Kliment Ohridski"
Partizanska bb
7000 Bitola
REPUBLIC OF NORTH MACEDONIA

E-mails: natasa.tabakovska@uklo.edu.mk
snezana.savoska@uklo.edu.mk
andrijana.bocevska@uklo.edu.mk
ilija.jolevski@uklo.edu.mk
blagoj.ristevski@uklo.edu.mk

Vladimir Trajkovik

University "Ss. Cyril and Methodius"
1000 Skopje
REPUBLIC OF NORTH MACEDONIA

E-mail: vladimir.trajkovik@finki.ukim.mk

ГОДИШНИК НА СОФИЙСКИЯ УНИВЕРСИТЕТ „СВ. КЛИМЕНТ ОХРИДСКИ“
ФАКУЛТЕТ ПО МАТЕМАТИКА И ИНФОРМАТИКА
Том 110

ANNUAL OF SOFIA UNIVERSITY “ST. KLIMENT OHRIDSKI”
FACULTY OF MATHEMATICS AND INFORMATICS
Volume 110

CAPEC ONTOLOGY

VLADIMIR DIMITROV

CAPEC is an effort coordinated by MITRE Corporation. Its aim is attack pattern database structured in taxonomies. CAPEC is available as XML document from its project site. CAPEC structure and content are under permanent change and development. It is still not mature database but may be never will.
CAPEC, CWE, and CVE are databases devoted to attacks, weaknesses, and vulnerabilities. They refer each other forming a knowledge ecosystem in cybersecurity area.
Traditional approach for knowledge presentation as information does not work well with conceptualizations under dynamics of this ecosystem and particularly of CAPEC. In this paper an alternative approach to CAPEC knowledge presentation is proposed, as ontology. First, CAPEC structure and content are discussed and then ontology structure is introduced. CAPEC as ontology opens doors to “open world” concept that is more adequate to ecosystem dynamics.
Keywords: cybersecurity, attack patterns, ontology, CAPEC, OWL
CCS Concepts:
• Security and privacy∼Formal methods and theory of security∼Formal security models;
• Security and privacy∼Formal methods and theory of security∼Logic and verification

1. Introduction

CAPEC (Common Pattern Enumeration and Classification) [2] is an effort coordinated by MITRE Corporation. Its aim is attack pattern database structured in taxonomies.

CAPEC is freely available in XML format from the site.

Some basic terms from [2]:

- “An attack pattern is a description of the common attributes and approaches employed by adversaries to exploit known weaknesses in cyber-enabled capabilities. Attack patterns define the challenges that an adversary may face and

DOI: 10.60063/GSU.FMI.110.63-83

how they go about solving it. They derive from the concept of design patterns applied in a destructive rather than constructive context and are generated from in-depth analysis of specific real-world exploit examples.”

- “An attack pattern is the common approach and attributes related to the exploitation of a weakness in a software, firmware, hardware, or service component.”
- “An attack (noun) is the use of an exploit(s) by an adversary to take advantage of a weakness(s) with the intent of achieving a negative technical impact(s). An attack is part of the bigger "Cyber Attack Lifecycle" that includes the following tasks: reconnaissance, weaponize, deliver, exploit, control, execute, and maintain.”
- “An exploit (noun) is an input or action designed to take advantage of a weakness (or multiple weaknesses) and achieve a negative technical impact. The existence (even if only theoretical) of an exploit is what makes a weakness a vulnerability.”

Attack pattern is an abstraction mechanism that describes execution of known attack. It is a blueprint for an exploit; has applicability context and recommended attack mitigation methods.

Attack pattern concept comes from design patterns [1] in object-oriented design, but in destructive context.

Design patterns are common design problem solutions at high level. They cannot be used directly for codding. Instead, design pattern is a description how the problem can be solved.

However, attack patterns are not very abstract. They are related to some functionality type and some exploited weaknesses.

Attack patterns are not very specific related to some concrete application.

Note 1. The term “structured text” below is plain text – literal structured in accordance with some rules. It does not contain special markers.

Note 2. CAPEC elements are defined with XSD types. It is more readable to present elements by names but not by types. In [2] schema is presented by types but only element names are visible in CAPEC catalog – not the types. That is the presentation way used in next section.

2. CAPEC structure

CAPEC catalog (Attack_Pattern_Catalog) contains elements Views, Categories, Attack_Patterns, and External_References. The last ones contain elements View, Category, Attack_Pattern, and External_Reference correspondingly.

Attack_Pattern_Catalog have attributes Name, Version, and Date – required.

2.1. CAPEC structure

View organizes attack patterns from specific point of view. It has attributes ID, Name, Type, and Status – required.

There are three type of views: Explicit, Implicit, and Graph. The catalog currently contains Implicit and Graph views.

Explicit view simply enumerates its members.

Implicit view dynamically defines its members via XPath query.

Graph view has hierarchical structure. Graph views have as children only categories in the current catalog, and the categories have only meta attacks. It is not controlled by CAPEC XSD schema.

CAPEC-533 is obviously an explicit view but is represented as implicit one – its members are simply enumerated.

The view contains elements discussed below.

Objective element is a required structured text describing the view perspective.

Audience element is a list of Stakeholder elements. Every Stakeholder element has two required element: Type and Description. Type element is enumeration with values: Academic Researchers, Applied Researchers, Assessment Customers, Assessment Vendors, CAPEC Team, Educators, Information Providers, Software Customers, Software Designers, Software Developers, Software Vendors, and Other. Description element is a structured text.

Filter element set XPath query for implicit views.

References element is sequence of Reference elements. These are external references.

Reference element has External_Reference_ID and Section attributes. First attribute is in REF-n format – something like footnote. Section points to specific location in the reference.

Notes element is list of Note elements. This is additional information of any kind.

Note element is structured text. Its Type attribute may have as value one of Maintenance, Relationship, Research Gap, Terminology or Other.

Content_History element contains catalog history. Its elements are Submission, Modification, Contribution, and Previous_Entry_Name.

Submission element has elements Submission_Name, Submission_Organization, Submission_Date, and Submission_Comment. Except Submission_Date that is date, all they are strings.

Modification element has elements Modification_Name, Modification_Organization, Modification_Date, Modification_Importance, and Modification_Comment. All they are strings except Modification_Importance that can has as value Normal or Critical.

Contribution element has elements Contribution_Name, Contribution_Organization, Contribution_Date, and Contribution_Comment. Its required attribute Type can have as value Content or Feedback.

Previous_Entry_Name element is a string with required attribute Date.

Members element links explicit and graph views with its members. In graph views, Members points only to categories. There are no explicit views in CAPEC catalog.

There is Relationships element in Categories with same type and purpose as Members in Views.

Members element may contain arbitrary number of Member_Of and Has_ Member elements. There are no restrictions on view members in CAPEC XSD schema, but in the catalog contents, graph views have categories for members and Relationships in categories – only attack patterns. Implicit views have as members attack patterns.

Members_Of and Has_Member elements have the same structure. CAPEC_ID attribute can point to any catalog entry by CAPEC XSD schema, but in practice, it is not true – it is discussed below.

Members_Of and Has_Member elements may have arbitrary number of Exclude_Related elements.

Exclude_Related element has Exclude_ID attribute. It is CAPEC identifier, i.e. view, category or attack pattern. The idea of this element is to exclude the ancestor from the relationship (Members_Of and Has_Member).

This mechanism has no sense for views and categories because there is no Members_Of / Has_Member three with path longer than one.

The same element is used in Related_Attack_Patterns element. It is discussed below, but situation there is more confusing.

In reality, CAPEC catalog does not contain Exclude_Related elements for categories and views. Something more, catalog contains only Has_Member elements.

2.2. Categories

Categories element contain Category elements.

Category element has required attributes ID, Name, and Status with the same meaning and purpose as that ones in views.

Summary is required structured text for categories.

Relationships element describes the abstractions hierarchy as Members element for views.

References, Notes, and Content_History have the same structure and meaning as that ones for views.

Taxonomy_Mappings element has Taxonomy_Mapping elements.

Taxonomy_Mapping element has required Taxonomy_Name attribute that can have as value ATTACK, WASC, or OWASP Attack.

Taxonomy_Mapping element subelements are Entry_ID, Entry_Name, and Mapping_Fit. They point to the target taxonomy element and describe mapping effectiveness with the last element that can as value Exact, CAPEC More Abstract, CAPEC More Specific, Imprecise, or Perspective.

2.3. Attack patterns

Attack_Pattern element has required attributes ID, Name, Status, and Abstraction. The last can have as values Meta, Standard, and Detailed.

Meta attacks stay at highest abstraction level. They are defined by specific methodology or attack technique but not by specific technology or implementation.

Meta attacks group standard attacks. They are useful in Architecture/Design phase.

Standard attacks are focused on specific attack technique or methodology. Frequently, they are part of "whole attack".

Standard attacks contain enough details to be understand their specific technique and how achieve its aim.

Detailed attacks are at the lowest abstraction level. They are focused on specific technique or technology. Detailed attack describes the whole execution flow.

Detailed attacks contain protection descriptions against them.

Very frequently, detailed attacks are chains of standard attacks. Therefore, detailed attacks can be placed at several locations in the attack taxonomy.

Description and Extended_Description elements are structured text.

Alternate_Terms element has Alternate_Term elements.

Alternate_Term element has required Term element and optional Description element that is structured text.

Likelihood_Of_Attack element is enumeration of High, Medium, Low, and Unknown values.

Typical_Severity element is enumeration of Very High, High, Medium, Low, and Very Low values.

Execution_Flow element has elements Attack_Step.

Attack_Step has required Step, Phase, Description, and optional Technique elements.

Step element numerates the step.

Phase element is enumeration of Explore, Experiment, and Exploit values.

Technique element is structured text extended with CAPEC_ID attribute. The last one can point to any catalog entry – there is no restriction in CAPEC XSD schema, but has to point to attack technique.

Prerequisites element has Prerequisite elements.

Prerequisite element is a structured text describing conditions for attack success.

Skills_Required element contains Skill elements.

Skill element is string extended with Level attribute that can have value High, Medium, Low, or Unknown.

Resources_Required element has Resource elements that are structured text.

Indicators element contains Indicator elements – structured text. Indicators are activities, events, conditions, or behaviors describing attack preparations, attack under development, or results of successful attack.

Consequences element contains Consequence elements.

Consequence element has Consequence_ID attribute and Scope, Impact, Likelihood, and Note elements.

Scope element is enumeration of Confidentiality, Integrity, Availability, Access Control, Accountability, Authentication, Authorization, Non-Repudiation, and Other.

Impact element is enumeration of Modify Data, Read Data, Unreliable Execution, Resource Consumption, Execute Unauthorized Commands, Gain Privileges, Bypass Protection Mechanism, Hide Activities, Alter Execution Logic, and Other.

Likelihood is enumeration of High, Medium, Low, and Unknown.

Note element is structured text.

Mitigations element has Mitigation elements – structured text describing how to harden the system, to shrink attack surface, or decrease impacts from successful attack.

Example_Instances element has Example elements – structured text.

Related_Weaknesses has Related_Weakness elements.

Related_Weakness element has required attribute CWE_ID – pointing to CWE weakness.

Taxonomy_Mappings, References, Notes and Content_History elements are the same as in categories.

Finally, Related_Attack_Patterns element has Related_Attack_Pattern elements.

Related_Attack_Pattern element has required Nature and CAPEC_ID attributes. CAPEC_ID can point to any catalog entry by CAPEC XSD schema – not only to attack patterns how the name suggests.

Nature attribute is enumeration of ChildOf, ParentOf, CanFollow, CanPrecede, CanAlsoBe, and PeerOf.

Pairs CanFollow/CanPrecede organize unnamed attack patterns.

CanAlsoBe link the attack with a similar one, but inverse relationship is not obligatory.

PeerOf marks relationships that cannot be classified as some of preceding ones.

Pair ChildOf/ParentOf sets relationships between attack patterns at different abstract levels, but there are no such restrictions in CAPEC XSD schema.

Only ChildOf relationship is used in the catalog.

There are only two graph views in the catalog. Meta attacks in them point to standard or detailed attacks.

Related_Attack_Pattern element can contain Exclude_Related elements.

Exclude_Related element has required Exclude_ID attribute that is CAPEC identifier. There are no restrictions in CAPEC XSD schema for Exclude_ID.

Following the logic of relationships ChildOf/ParentOf, only attack patterns participating in this kind of relationships have to be excluded.

CAPEC catalog contains only ChildOf relationships. ChildOf/ParentOf relationship trees start from meta attacks. Therefore, exclusion of an ancestor means to prune the subtrees starting from this ancestor.

An attack pattern can participate in several subtrees of taxonomy hierarchy.

Following this logic, only attack patterns have to be excluded, but in the catalog, only categories are excluded.

Categories and attack patterns can be in Member_Of/ Has_Member relationships. Therefore, ChildOf/ParentOf relationship is extended with Member_Of/ Has_Member relationship.

Something more, there are views (not the basic ones) in the catalog, in which categories directly point to standard attacks without meta attacks.

2.4. External references

External_References elements are sequence of External_Reference elements.

External_Reference element has required Reference_ID attribute and elements Author, Title, Edition, Publication, Publication_Year, Publication_Month, Publication_Day, Publisher, URL, and URL_Date. It points to external detailed information.

3. CAPEC ontology

CAPEC ontology follows CAPEC XSD schema. Description data are annotations in this ontology. Classification data are classes, object, and data properties.

Constructions that have no clear semantic are more difficult for interpretation. In that case, thorough investigation of their real usage in the database has been done. Such examples are attack relationships.

Attack concept presentation is still immature that is why there are many constructs with unclear semantic. Therefore, the best approach is OWL and "open world" assumption for knowledge presentation.

CAPEC ontology includes as annotations some referent information: schema, date, copyright, dictionary name, version, version date, ontology author, and ontology version.

CAPEC ontology contains external references. They are presented as External_Reference annotations that are literals structured by Author:, Title:, Edition:, Publication year:, Publication month:, Publication day:, Publisher:, URL:, URL date: and Reference_ID: – subelements of External_Reference element.

Figure 1 shows CAPEC ontology common class schema.

Figure 2 shows CAPEC class hierarchy.

3.1. Views

CAPEC ontology is a set of views. View class presents view concept. This class is union without intersection of its subclasses Explicit, Graph, and Implicit.

Type data property presents view type and can has as value one of the abovementioned subclasses names. This is some unnecessary information duplication but for CAPEC XSD schema "compatibility" is available.

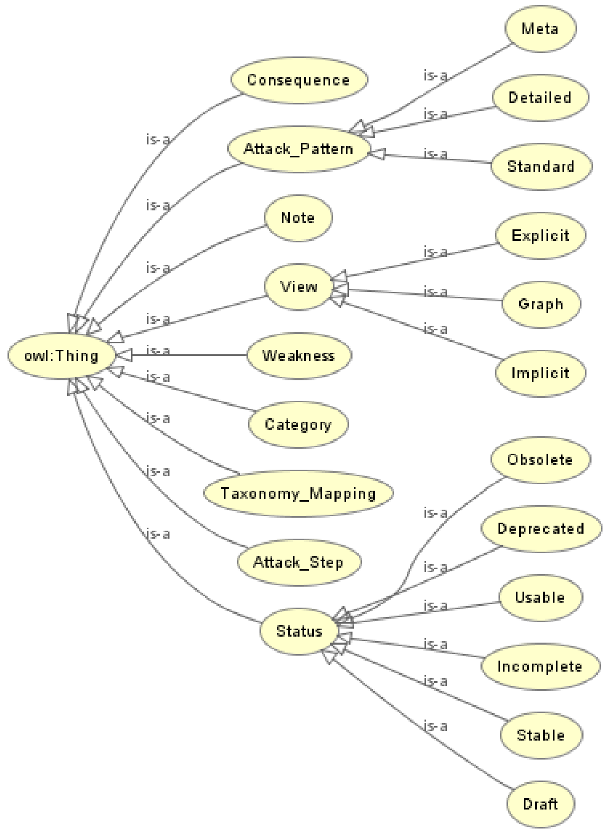

Figure 1. CAPEC ontology common class schema

Explicit class has no instances at all in the catalog. There are some explicit views defined as implicit ones – query of these implicit views simply lists class members.

There are two main graph views. They have strong hierarchical structure view – category – meta attack – standard attack – detailed attack. However, there are some deviations from this structure in other view – for example, some meta attacks may have as children detailed attacks.

Query defines implicit views. Reasoners do not execute queries. That is why implicit view content is deployed with Has_Member object property pointing to attack from query execution result.

```
Class: View
    DisjointUnionOf:
        Explicit, Graph, Implicit
Class: Explicit
    SubClassOf:
```

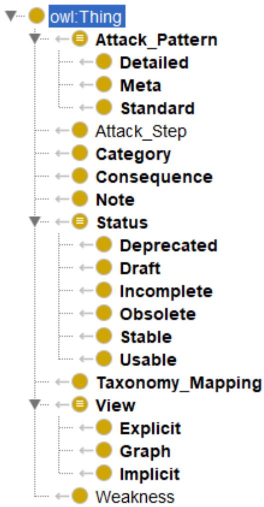

Figure 2. CAPEC class hierarchy

```
        View
Class: Graph
    SubClassOf:
        View
Class: Implicit
    SubClassOf:
        View
```

Annotation definitions will be not presented here.

View status is presented by Status class (categories and attacks have status too). Status class is union without intersection of views, categories, and attacks.

On the other hand, Status class has subclasses Deprecated, Draft, Incomplete, Obsolete, Stable, and Usable. Status class is union without intersection of these subclasses too.

Status class is not "abstract class" – class without instances. If an individual is from only Status class this means that still is not clear to which of Status classes this individual belongs.

Abstract classes can be controlled via SHACL [3] because Semantic Web uses "open world" assumption.

Above considerations are true for views.

```
Class: Status
    DisjointUnionOf:
        Attack_Pattern, Category, View
    DisjointUnionOf:
        Deprecated, Draft, Incomplete, Obsolete, Stable,
```

```
        Usable
Class: Deprecated
    SubClassOf:
        Status
Class: Draft
    SubClassOf:
        Status
Class: Incomplete
    SubClassOf:
        Status
Class: Obsolete
    SubClassOf:
        Status
Class: Stable
    SubClassOf:
        Status
Class: Usable
    SubClassOf:
        Status
```

Inverse pair of object properties Has_Member/Member_Of presents membership of categories and attacks to view. CAPEC XSD schema does not control in any way this membership. It is possible some view to be member of another view.

Total control on memberships in the ontology can be achieved via SHACL.

Above consideration are applicable for category membership.

Following the membership relationships, Has_Member object property has as domain views and categories, and as range views, categories, and attacks.

```
ObjectProperty: Has_Member
    Domain:
        Category or View
    Range:
        Attack_Pattern or Category or View
    InverseOf:
        Member_Of
ObjectProperty: Member_Of
    Domain:
        Attack_Pattern or Category or View
    Range:
        Category or View
    InverseOf:
        Has_Member
```

Objective element is presented as annotation.

View has Audience data property of enumeration type of values Academic Researchers, Applied Researchers, Assessment Customers, Assessment Vendors, CAPEC

Team, Educators, Information Providers, Other, Software Customers, Software Designers, Software Developers, and Software Vendors.

```
DataProperty: Audience
    Domain:
        View
    Range:
        {"Academic Researchers",
         "Applied Researchers",
         "Assessment Customers", "Assessment Vendors",
         "CAPEC Team", "Educators", "Information Providers",
         "Other", "Software Customers",
         "Software Designers", "Software Developers",
         "Software Vendors"}
```

Audience data property can be accompanied by Audience_Description annotation.

Audience data property and Audience_Description annotation come from Stakeholder element. The last is not used in the ontology.

Only implicit views have Filter annotation. It contains XPath query from which view content is deployed via Has_Member object property.

Reference annotation presents view references – an annotation for a reference. Here, External_Reference_ID and Section elements are embedded on separate lines.

View notes have structure that is more informative. They are deployed via Note class and Note object property pointing to individuals of that class. Notes have Type data property and Note_Description annotation. Note's Type can have as value Maintenance, Relationship, Research Gap, Terminology, or Other.

```
ObjectProperty: Note
    Domain:
        Attack_Pattern or Category or View
    Range:
        Note
Class: Note
DataProperty: Type
    Characteristics:
        Functional
    Domain:
        Note
    Range:
        {"Maintenance", "Other", "Relationship",
         "Research Gap", "Terminology"}
```

Current_History annotation presents the catalog element with the same name. Its subelements (Submission, Modification, Contribution, and Previous_Entry_Name) are embedded in Current_History on separate lines with offsets.

References, notes, and content history are available for categories and attacks with the same structure, interpretation, and meaning.

View individual IRI is constructed with view ID attribute. The last is ID number prefixed with "CAPEC_". Category and attack individuals construct their individual IRIs in the same way.

```
DataProperty: ID
    Characteristics:
        Functional
    Domain:
        Attack_Pattern or Category or View
    Range:
        xsd:positiveInteger
```

View name is presented with Name data property. Category and attack names are presented with the same data property.

```
DataProperty: Name
    Characteristics:
        Functional
    Domain:
        Attack_Pattern or Category or View
    Range:
        xsd:string
```

3.2. Categories

Category class presents categories. Category groups attacks on particular target or effect.

```
Class: Category
```

Category description aim is given with Summary annotation.

Category members are linked with its category via Member_Of object property (inverse object property is Has_Member). These object properties are described in view subsection.

Categories can be mapped to other taxonomies – not only which are presented in CAPEC. This mapping is more informative. Taxonomy_Mapping class, and object property with the same name are used for that purpose. The object property points to individuals from the class.

```
Class: Taxonomy\_Mapping
```

Taxonomy_Mapping class has Taxonomy_Name data property. The last may have as value ATTACK, WASC or OWASP Attacks. Taxonomy_name is required but this can be controlled with SHACL.

```
DataProperty: TaxonomyName
    Characteristics:
        Functional
    Domain:
        Taxonomy_Mapping
    Range:
        {"ATTACK", "OWASP Attacks", "WASC"}
```

Entry_ID, Entry_Name, and Mapping_Fit fix the mapping. First two sets target taxonomy name and entry name in this taxonomy. Mapping_Fit describes how precise is the mapping: Exact, CAPEC More Abstract, CAPEC More Specific, Imprecise, or Perspective.

```
DataProperty: Entry_ID
    Characteristics:
        Functional
    Domain:
        Taxonomy_Mapping
    Range:
        xsd:string
DataProperty: Entry_Name
    Characteristics:
        Functional
    Domain:
        Taxonomy_Mapping
    Range:
        xsd:string
DataProperty: Mapping_Fit
    Characteristics:
        Functional
    Domain:
        Taxonomy_Mapping
    Range:
        {"CAPEC More Abstract", "CAPEC More Specific",
         "Exact", "Imprecise", "Perspective"}
```

3.3. Attack patterns

Attack pattern by its abstraction level can be meta, standard, or detailed. Attack_Pattern class and its subclasses Meta, Standard, and Detailed present attack patterns and their abstraction levels.

Attack_Pattern class is union without intersection of its subclasses.

Attack_Pattern class can be viewed as "abstract class" in the "closed world".

```
Class: Attack_Pattern
    DisjointUnionOf:
```

```
        Detailed, Meta, Standard
Class: Detailed
    SubClassOf:
        Attack_Pattern
Class: Meta
    SubClassOf:
        Attack_Pattern
Class: Standard
    SubClassOf:
        Attack_Pattern
```

References, notes, and content history are as in categories.

Attack_Pattern IRIs, statuses, and attack names are as in categories.

Attack_Pattern_Description annotation describes the attack. Required annotation eventually can be supported by SHACL.

Extended_Description annotation can be attached to attacks.

Alternative attack names are presented with Alternate_Term data property and its annotation Alternate_Term_Description.

```
DataProperty: Alternate_Term
    Domain:
        Attack_Pattern
    Range:
        xsd:string
```

Likelihood_Of_Attack is functional data property that can have one of High, Medium, Low, or Unknown.

```
DataProperty: Likelihood_Of_Attack
    Characteristics:
        Functional
    Domain:
        Attack_Pattern
    Range:
        {"High" , "Low" , "Medium" , "Unknown"}
```

Typical_Severity is functional data property that can have one of Very High, High, Medium, Low, or Very Low.

```
DataProperty: Typical_Severity
    Characteristics:
        Functional
    Domain:
        Attack_Pattern
    Range:
        {"High" , "Low" , "Medium" , "Very High" ,
         "Very Low"}
```

Execution_Flow object property points to individuals of Attack_Step class. Attack_Step individual IRI is constructed by attack's IRI suffixed with "_Attack_Step" and step number starting from zero.

```
ObjectProperty: Execution_Flow
    Domain:
        Attack_Pattern
    Range:
        Attack_Step
Class: Attack\_Step
```

Attack_Step class has required functional property Step. It is step sequence number. Numbering starts with one and has no relations with IRI individual construction for the class.

```
DataProperty: Step
    Characteristics:
        Functional
    Domain:
        Attack_Step
    Range:
        xsd:positiveInteger
```

Phase is required functional property of Attack_Step. It can have as value one of Explore, Experiment, and Exploit – attack phase in which the step is executed.

```
DataProperty: Phase
    Characteristics:
        Functional
    Domain:
        Attack_Step
    Range:
        {"Experiment" , "Exploit" , "Explore"}
```

Attack_Step individuals must be annotated with Attack_Step_Description.

Technique applicable in the attack step can optionally be another attack pattern. This option can be implemented via Technique object property pointing to another attack pattern. Nevertheless, CAPEC XSD schema permits this element to point to views and categories.

```
ObjectProperty: Technique
    Domain:
        Attack_Step
    Range:
        Attack_Pattern
```

If attack pattern is attached to the step then the object property may be annotated with Technique_Description. Otherwise, this annotation is applied to the step individual. Therefore, technique description can appear in two different locations. The idea may be is that in thoroughly analyzed systems, attack steps must be attack patterns.

Technique description appears in Technique element whose content is structured text optionally extended with CAPEC_ID attribute. The last sets the value of Technique object property.

```
ObjectProperty: CAPEC_ID
    Characteristics:
        Functional
    Range:
        Attack_Pattern
```

Prerequisite annotation presents preliminary requirements for attack success.

Required attacker skills for attack execution are presented with Skill data property that can have as value one of High, Medium, Low, or Unknown. This property can be annotated with Skill_Description.

```
DataProperty: Skill
    Domain:
        Attack_Pattern
    Range:
        {"High", "Low", "Medium", "Unknown"}
```

Required resources for successful attack are defined with Resource annotations.

Indicator annotations describe attack preparation, ongoing attack or finished attack indicators.

Attack consequences are described via Consequence object property pointing to individuals of class with the same name. Individual IRI of that class are constructed by attack IRI suffixed with "_Consequence" and consequence successive number starting from zero.

```
ObjectProperty: Consequence
    Domain:
        Attack_Pattern
    Range:
        Consequence
Class: Consequence
```

Consequence class has Scope data property with possible values Confidentiality, Integrity, Availability, Access Control, Accountability, Authentication, Authorization, Non-Repudiation, and Other, representing security areas impacted by successful attack execution. At least one security area must be marked.

```
DataProperty: Scope
    Domain:
        Consequence
    Range:
        {"Access Control", "Accountability",
         "Authentication", "Authorization", "Availability",
         "Confidentiality", "Integrity", "Non-Repudiation",
         "Other"}
```

Another data property of Consequence class is Impact with possible values Modify Data, Read Data, Unreliable Execution, Resource Consumption, Execute Unauthorized Commands, Gain Privileges, Bypass Protection Mechanism, Hide Activities, Alter Execution Logic, or Other. These are technical consequences of successful attack.

```
DataProperty: Impact
    Domain:
        Consequence
    Range:
        {"Alter Execution Logic",
         "Bypass Protection Mechanism",
         "Execute Unauthorized Commands", "Gain Privileges",
         "Hide Activities", "Modify Data", "Other",
         "Read Data", "Resource Consumption",
         "Unreliable Execution"}
```

Consequence likelihood is given by Likelihood data property, which is functional and can have as value one of High, Medium, Low, or Unknown.

```
DataProperty: Likelihood
    Characteristics:
        Functional
    Domain:
        Consequence
    Range:
        {"High", "Low", "Medium", "Unknown"}
```

Consequence may be attached with Consequence_ID identifier (for internal use) and Note annotation.

```
DataProperty: Consequence_ID
    Characteristics:
        Functional
    Domain:
        Consequence
    Range:
        xsd:string
```

Attack mitigations (preventions) are not formalized and are presented with Mitigation annotations.

Attack examples are not formalized too and are presented with Example annotations.

Weaknesses that are exploited by the attack are referred via Related_Weakness object property. There are no comments on weakness role. The property simply point to an individual from CWE ontology.

```
ObjectProperty: Related_Weakness
    Domain:
        Attack_Pattern
    Range:
        cwe:Weakness
```

Finally, related attack patterns are discussed. They generically are presented with Related_Attack_Pattern object property. This property has to be “abstract” in the “closed world”. “Abstract” classes have no instances, so “abstract” properties have no links (instances).

```
ObjectProperty: Related_Attack_Pattern
    Domain:
        Attack_Pattern
    Range:
        Attack_Pattern
```

Related_Attack_Pattern has subproperties ChildOf, ParentOf, CanFollow, CanPrecede, CanAlsoBe, and PeerOf. Each of them presents some kind of relationship.

Inverse pairs are ChildOf/ParentOf and CanFollow/CanPrecede.

```
ObjectProperty: ChildOf
    SubPropertyOf:
        Related_Attack_Pattern
    InverseOf:
        ParentOf
ObjectProperty: ParentOf
    SubPropertyOf:
        Related_Attack_Pattern
    InverseOf:
        ChildOf
ObjectProperty: CanFollow
    SubPropertyOf:
        Related_Attack_Pattern
    InverseOf:
        CanPrecede
ObjectProperty: CanPrecede
```

```
    SubPropertyOf:
        Related_Attack_Pattern
    InverseOf:
        CanFollow
ObjectProperty: CanAlsoBe
    SubPropertyOf:
        Related_Attack_Pattern
ObjectProperty: PeerOf
    SubPropertyOf:
        Related_Attack_Pattern
```

Inverse relationships form chains of relationships. ChildOf/ParentOf organize abstraction hierarchy. ChildOf, must link in principle the attack with another attack from the same or at higher abstraction level. However, there is no such restriction in CAPEC XSD schema.

De facto, CAPEC XSD schema does not put any restrictions on CAPEC_ID attribute of Related_Attack_Pattern element. It is possible this attribute to point even to view or category.

Obviously, authors of CAPEC XSD schema have not fixed above-mentioned problem and Related_Attack_Pattern intend has to be detected from the database content. It is possible in future the problem to be fixed and even sustainable changes to CAPEC XSD schema to be done.

In CAPEC database content, Related_Attack_Pattern points only to attack patterns. It is possible to use SHACL for more strict control on these relationships, for example, ChildOf to point to attack patterns on the same or higher abstraction level.

Pair CanFollow/CanPrecede may organize chains of unnamed attack patterns. Attack chain concept is similar to that one in CWE but not so advanced.

CanAlsoBe and PeerOf do not organize relationship chains. CanAlsoBe relates pair of attacks but this relationship is not commutative or transitive. PeerOf relates two attacks in relationship that cannot be classified as relationship from the other five listed kinds.

Related_Attack_Pattern property has for domain only attack patterns.

From analysis of database's content follows that follows that only categories are excluded from relationship trees. Therefore, exclusion is not applied as universally as defined by CAPEC XSD schema. In conclusion, abstraction subtrees organized by ChildOf/ParentOf and Member_Of/Has_Member are pruned only on category level.

It is desirable CAPEC XSD schema to be changed following the real construction usage. Even more, many more upgrades of this schema are expected in the future. For now, CAPEC ontology follows current CAPEC XSD schema and some restrictions are applied considering CAPEC database content.

Now ancestor exclusion will be discussed. From content of CAPEC database follows that only category ancestors are excluded. Between excluded category and the attack under consideration there are zero or more attack linked with that category. The last can be at any abstraction level by CAPEC XSD schema, but in practice, levels are meta and standard. Excluded category is linked with attack under consideration via Member_Of/Has_Member and zero or more ChildOf/ParentOf relationships. Ancestor tree starts from view, then pass to categories via Member_Of/Has_Member, and after that via zero or more ChildOf / ParentOf ends in the attack under consideration.

Category children define category of its descendants in ancestor tree – some kind of category type. Therefore, via Exclude_Related elements only specific categories are excluded as ancestors for the attack under consideration. It follows that independent of the fact that there is a path organized by Member_Of/Has_Member and ChildOf/ParentOf, only categories are excluded. As result of these considerations Exclude_Related object property excludes only categories. May be in the future this restriction will be changed – maybe not, or CAPEC taxonomy or CAPEC XSD schema will be changed.

4. Conclusion

CAPEC schema analysis and especially analysis on relationships shows that CAPEC is still not mature. Therefore, CAPEC schema upgrades will be developed in the future.

In this situation, is there any sense to formalize CAPEC knowledge in OWL? Concept of "closed world" supposes that all knowledge is collected in knowledge base. This means that statements not following from this knowledge base are not true.

In "open world" concept if something is not known may be true.

In mature taxonomy, the database contains all knowledge. This means that when CAPEC get mature, its database would cover "closed world" criteria. Is it true?

Currently, there are very many questions about some attacks, i.e. this problem area is under intensive investigations and is way from "closed world".

On the other hand is the future. Eventually, CAPEC taxonomy will get stable enough to classify it as mature. The problem to happen this is that new complicated unknown attacks are under development. Therefore, there is no way CAPEC taxonomy to get mature. In this situation more applicable is "open world" concept, because CAPEC taxonomy is unstable – under permanent development.

What is NIST approach to CVE? NIST extracts from CVE only analyzed vulnerabilities for NVD. In NVD, NIST guarantees the "closed world" concept.

Why CPE, NVD, CVE, CWE, and CAPEC are not presented as ontologies? All these databases are under permanent development. If these databases were presented as ontologies, there would be no need to do transformations like CVE to NVD. Something more, reasoners can open new horizons for research and investigations. Finally, "closed world" aspects can be checked with SHACL.

Acknowledgements

This paper is prepared with the support of MIRACle: Mechatronics, Innovation, Robotics, Automation, Clean Technologies – Establishment and development of a Center for Competence in Mechatronics and Clean Technologies – Laboratory Intelligent Urban Environment, funded by the Operational Program Science and Education for Smart Growth 2014–2020, Project BG 05M2OP001-1.002-0011.

References

[1] E. Gamma, R. Helm, R. Johnson, J. Vlissides, Design patterns: Elements of reusable object-oriented software, Boston, MA: Addison-Wesley, 1995.

[2] MITRE Corporation, CAPEC (Common Pattern Enumeration and Classification), 2023. URL: http://capec.mitre.org.

[3] W3C, Shapes Constraint Language (SHACL), W3C Recommendation 20 July 2017. URL: http://www.w3.org/TR/shacl.

Received on March 20, 2023
Accepted on May 7, 2023

Vladimir Dimitrov

Faculty of Mathematics and Informatics
Sofia University "St. Kliment Ohridski"
5 James Bourchier Blvd.
1164 Sofia
BULGARIA

E-mail: cht@fmi.uni-sofia.bg

ГОДИШНИК НА СОФИЙСКИЯ УНИВЕРСИТЕТ „СВ. КЛИМЕНТ ОХРИДСКИ“
ФАКУЛТЕТ ПО МАТЕМАТИКА И ИНФОРМАТИКА
Том 110

ANNUAL OF SOFIA UNIVERSITY “ST. KLIMENT OHRIDSKI”
FACULTY OF MATHEMATICS AND INFORMATICS
Volume 110

MACHINE LEARNING FOR BIOMEDICAL-FOCUS ON BRAIN SIGNALS

MOHAMAD FAKHREDINE and TEODORA BAKARDJIEVA

There is a growing interest in machine learning (ML) in this decade. This growing interest is accelerated by cheaper computing power and low-cost memory. Thus, large amount of data can be stored, processed and analyzed efficiently. Machine learning has used in brain machine systems (BMS) system that converts brain impulses into messages or commands. In this paper we propose an EEG-based BMS with the focus on evoked potential. An average classification accuracy of 95% was attained among nine participants. With a rate of 4 flashes per second implemented, selecting one of four possibilities takes 5 s, resulting in an information transfer rate of 24 bits/min. Also, brain computer interfacing using oscillatory activity was measured. The results show that after around 5 h of co-adaptive training over many days, the average 3-class accuracy of the Linear Discriminant Analysis committee classifier reached about 80%, with a false positive rate for motor imagery recognition of around 17%.

Keywords: brain-machine system, EEG, machine learning, classification, accuracy, false alarm rate, human-centered computing

CCS Concepts:
• Computing methodologies∼Machine learning∼Machine learning approaches;
• Human-centered computing∼Human computer interaction∼HCI theory, concepts and models

1. Introduction

During the last decade, the world has witnessed a significant technological breakthrough in Artificial Intelligence (AI). Artificial intelligence (AI) is defined as a branch of science and engineering concerned with the computational understanding of what is commonly referred to as intelligent behavior. Artificial intelligence has been extensively used in recent decades to build applications in fields like deep

DOI: 10.60063/GSU.FMI.110.85-94

learning, and machine learning AI is becoming a well-known in computer science field as it has improved human life in numerous ways. AI has recently outperformed humans across several domains, and there is big hope that this will happen in healthcare [7]. The field of computer vision has made incredible strides thanks to recent developments in deep learning techniques, which were made possible by improved computing power and the availability of large data sets [1].

Modern assistive technology can help people with severe motor impairments communicate better, manage their home environments, and move around more easily depending on their remaining motor abilities. Systems known as Brain-Machine Systems (Figure 1) are able to translate brain activity into signals that can command external machinery. They may therefore be the only means for people who are severely disabled to maintain or increase their communication and control options [2].

Bypassing the conventional neuro-muscular output routes, BMS use machine learning and digital signal processing to transform brain signals into activity. These systems were initially intended for biomedical applications, such as regaining lost motor function and enabling communication in patients who are completely paralyzed. BMSs have been investigated by more target markets as a novel input device for entertainment and gaming geared for non-disabled persons. It is the responsibility of the BMS to identify and anticipate behaviorally motivated changes in a user's brain signals, also referred to as "cognitive states."

On the other hand, it is well known now that brain signals can be captured non-invasively using electrodes on the scalp (electroencephalogram, EEG). EEG recording techniques and signals have been extensively used to establish techniques to help and impaired people to control a range of devices using BMS.

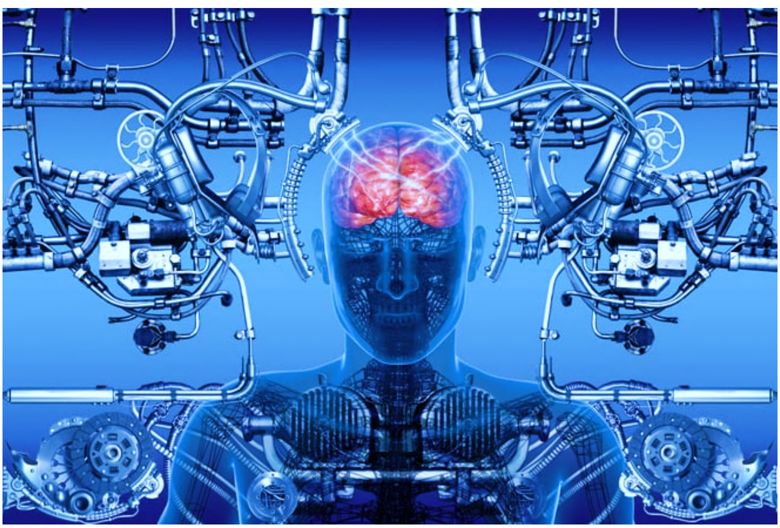

Figure 1. Brain-Computer Interfacing

The user must produce various brain activity patterns that the BMS will recognize and translate into commands in order to operate it. This identification is predicated on a classification technique used in the majority of current BMS. There have been a number of highly interesting BMS reviews published up to this point, but none have been devoted exclusively to a review of the classification algorithms used for BMS, their characteristics, and their assessment [3].

Specifically, chemicals, cells, vasculature, physical morphology, and cognition are all impacted by aging in the human body and brain. As time goes on, biological aging causes dependency and disability. The issue of aging and its effects on daily activities has been addressed in a number of diverse approaches, according to academics. Even for elderly patients and healthy older people, maintaining balance while talking, walking, and paying attention might be difficult. These limitations may make it difficult to carry on family conversations, navigate stairs, retain new knowledge, or operate a vehicle safely. Even if it's just utilizing handrails or canes to climb stairs, most elderly citizens need assistive equipment to help them do daily tasks, despite the fact that everyone's aging process has a varied impact. Unfortunately, they are not given the assistance they require since the expense of caring for them is in the billions of dollars.

BMS technology is increasingly being used to treat many individuals who have cognitive or physical problems. This system has the potential to significantly improve these patients' quality of life by increasing their personal autonomy and mobility. This paper proposes a novel implementation of classification algorithms used in EEG-based BMS applications for the objective of testing these algorithms based on evoked potentials and their accuracy level.

2. Literature review and background

The electroencephalogram (EEG) is a type of brain scan that measures electrical activity in the brain, according to study (Figure 2) [4]. It was written by Avid and published in the article EEG Signal Processing for BMS Applications. By placing electrodes on the scalp and using conductive paste, it is possible to record and analyze the brain's electrical signal, which is measured in microvolts [5]. These evoked potentials are a neurophysiologic study that analyzes the involvement of the auditory sensory system, visual system, and somatosensory pathways through evoked responses to a recognized and standardized stimulus in EEG data. Some of the several kinds of event-related evoked potentials (ERP) and visual evoked potentials (VEP) include auditory evoked potentials (PEA), motor evoked potentials (MRP), and steady state visual evoked potentials (SSVEP).

The goal of a BMS is to recognize and estimate behaviorally induced changes in a user's brain signals, also known as "cognitive states." BMSs built on these recording methods have given both healthy and disabled people the ability to operate a variety of appliances. The BMS system design should be dictated by the characteristics of the EEG feature(s), as this will determine the most efficient system for a specific user. The users are currently tested using the SMR- and P300-based BMS systems, and

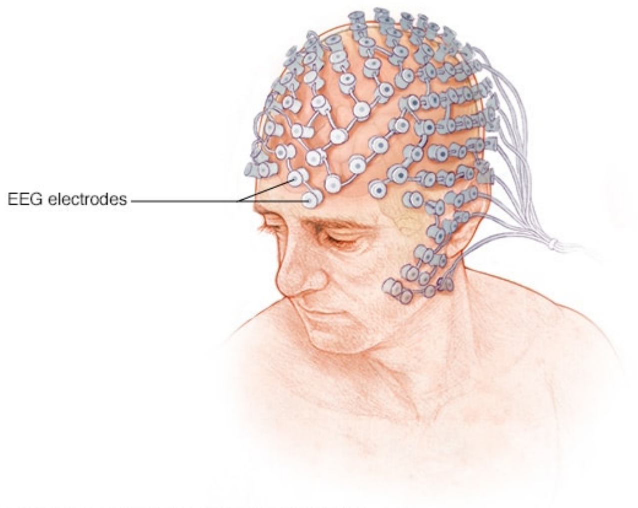

Figure 2. EEG System

the best system is chosen after consideration of speed, accuracy, bit rate, usefulness, and likelihood of use [6]. As we introduce BMS into people's homes, this model may prove to be the most effective.

According to [4], signal processing in the brain occurs in four steps: data acquisition, where brain signals are first acquired using sensors; pre-processing, where information that reduces signal quality is referred to as "noise" and "artifacts"; pre-processing techniques like spatial and temporal filtering can significantly improve signal quality by minimizing noise and artifacts; and feature extraction, where from the pre-processed brain signals, useful signal features are extracted. The majority of brain patterns used in BMS, such as slow cortical potentials (SCP), event-related potentials (ERPs), event-related desynchronization (ERD), and event-related synchronization (ERS), are mirrored by mu and beta rhythms (ERS). ERPs include the P300 and the steady-state visual evoked potential (SSVEP). Finally, the collected features are converted into discrete or continuous control signals that run devices such as a computer, wheelchair, or prosthesis using translation algorithms.

The author of [9] demonstrates how several neurons contribute to the signal measured at a particular electrode, comparing brain signals to the "cocktail party phenomena." The signals from these neurons are integrated and then aggregated at the electrode. Contrary to PCA, which maximizes the variance of the first component in the transformed space, common spatial patterns (CSP) optimize the variance-ratio of two conditions or classes. To put it another way, CSP looks for a transformation that maximizes the variance of samples from one condition while minimizes the variance of samples from the other condition. Temporal filtering, also known as frequency or spectrum filtering, is a crucial component of BMS systems for enhancing signal-to-noise ratio. Temporal filtering, for example, can be used to remove 50 Hz (60 Hz) noise from power sources. Temporal filtering can also be used to isolate motor imagery components from brain signals in a certain frequency band, such as the mu or beta band. One of the neurophysiological signals that can be used with BMSs is SSVEP [10].

The two primary EEG data types employed in BMSs are oscillatory activity patterns and evoked potentials (EPs). When a rare visual input or other external sensory event occurs, EPs, which are phase-locked electrical potential alterations, happen. EEG data are often averaged over time, commencing at the onset of the sensory event and lasting up to one second, in order to study EPs. On the other hand, oscillatory activity patterns can be purposefully created by the user. Most of the time, such imaging results in a power change in particular frequency bands [11].

The user may only need to make a small adjustment because EPs are repetitive brain reactions that are consistent over time. But as a result of feedback from BMS use, a user's oscillatory patterns frequently alter over time, making off-line learning of parameters less effective. In order to observe changes in oscillatory patterns and update the BMS as necessary, brain signals recorded during feedback are examined. The two examples below demonstrate the use of EPs and oscillatory patterns to connect the brain to the computer both physically and virtually [13].

An SVM also uses a discriminant hyperplane to define classes. However, the selected hyperplane in the case of SVM maximizes the margins. Margin maximization is known to enhance generalization skills. Using linear decision boundaries for classification is possible with linear SVM. The usage of this classifier in a sizable variety of synchronous BMS scenarios has been successful. However, with only a slight increase in the classifier's complexity, it is possible to set up nonlinear decision boundaries using the "kernel technique." It includes using a kernel function to indirectly transfer the data to a different space, typically one with much greater dimensions [12].

Furthermore, linear discrimination has been used in BMS as a technique for transforming high-dimensional feature data into a lower-dimensional space. It is then easier to divide the predicted data into two groups. The reparability of data is quantified in Fisher linear discriminant (FLD) by two quantities: the distance between projected class means (which should be large) and the amount of the data variance in this direction (should be small) [14]. The goal of LDA (also known as Fisher's LDA) is to segregate data representing various classes using hyperplanes. LDA presupposes that the data has a normal distribution and an equal covariance matrix for both classes. The separation hyperplane is found by locating the projection that maximizes the distance between the means of the two classes while minimizing the inter-class variation [16].

In fact, not all electrodes placed throughout the entire head are suitable in BM systems. As a result, channel selection is used to choose the most usable channels based on the attributes of a BM system. The EEG signals from the visual cortex are chosen for SSVEP-based BMSs. The channels in the sensorimotor area are chosen for mu/beta based BMSs. The channels with an obvious P300 are chosen for P300-based BMSs. In BMS, temporal filtering is critical for enhancing the signal-to-noise ratio, and it is also useful for extracting band power information.

3. Study design and methodology

A BMS enables a person to communicate with or control a device such as a computer or prosthetic without the use of peripheral nerves or muscles. A BMS in general architecture, which involves six stages of brain signal processing is shown in Figure 3. This procedure is generally separated into six parts, which are as follows: signal capture, preprocessing, feature extraction, classification, translation and feedback. The basic step involves skin preparation and electrode gel administration in the case of an EEG-based BMS. It is required because, large and unequal electrode impedances can degrade signal quality. Good signal quality, on the other hand, is critical for making the already tough task of retrieving information from brain signals easier [16]. Any BMS must be able to convert brain impulses into messages or commands. For signal translation, many signal processing and machine learning approaches have been developed. Most BMS systems require preprocessing because a good preprocessing method enhances the signal-to-noise ratio and spatial resolution, which improves the performance of the BMS. Preprocessing the EEG brain signal is a crucial step in the BMS because it creates an effective signal for the learning stages of detection and classification. From the pre-processed brain signals, useful signal features indicating the user's intent are retrieved. The classification algorithm is tuned after feature extraction, and the best features from numerous EEG channels are chosen and classified. Mainly, preprocessing, feature extraction, and classification are the three main steps. Following the extraction of features that reflect the user's intents, the next step is to transform these discriminative features into commands that run a device. These commands can be discrete in nature, such as letter selection, or continuous in nature, such as pointer movement vertically and horizontally. In BMSs, the best frequency bands are usually subject-specific. A

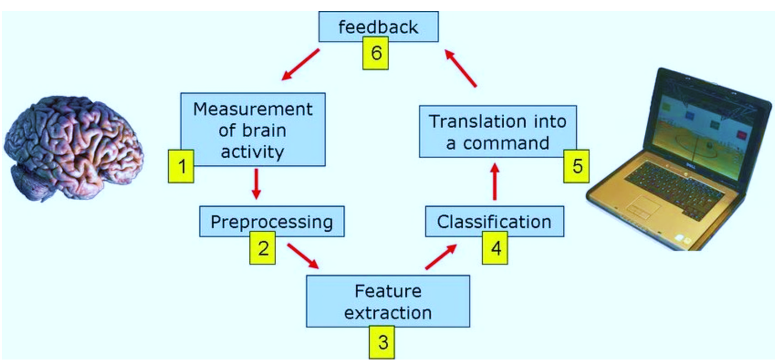

Figure 3. BMS architecture

feature selection method is frequently used to choose frequency bands. A wide frequency band is initially separated into multiple sub-bands that may overlap to carry out frequency band selection in this manner. Finally, feedback is conducted.

The parts of brain signals that reflect the user's objectives are useful, whereas the rest is likely noise. The P300 potential, for example, is a beneficial component in a P300-based BMS speller, whereas noise components include high frequency signals, alpha waves, EMG, EOG, and interference from power supplies and other equipment. The channels with an obvious P300 are chosen for P300-based BMSs.

4. Results and discussion

4.1. Brain machine system based on using evoked potentials

Figure 4 depicts an evaluation and monitoring in which the P300 is used to allow a user to choose from a menu of options. On a computer screen, the options are shown in a grid arrangement. The options in this experiment correspond to segmented images of items from a helper robot's current range of vision. The P300 (Figure 4A, panel 1) is utilized to figure out which object the BMS user wants the robot to pick up and carry to another area (Figure 4A, panels 2–4). The menu changes to image of possible destination locations once the object has been picked up, and P300 is used again to infer the user choice of desired robot's location [15].

The user concentrates his or her attention on the image of choice while the borders of the images are flashed one at a time in random order, to make a selection with the P300. That to say each image is flashed multiple of times in this random

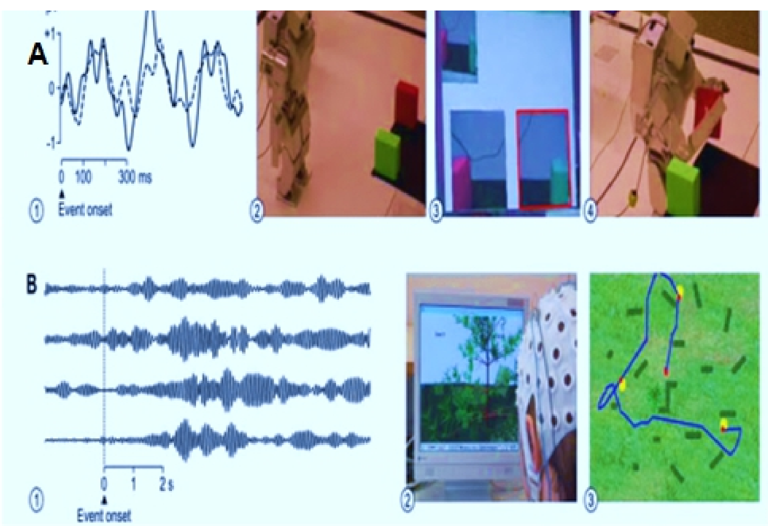

Figure 4. BMS system implementation

order, thus flashed on chosen image create P300 responses while other flashes do not.

A low-dimensional feature vector input was produced by applying a modest set of spatial filters to 32 channels of EEG data acquired from electrodes placed throughout the entire scalp as the SVM's input. These filters are called "spatial" because they are applied to the 32 samples that are spatially scattered throughout the scalp, rather than to samples over time. At each time step, the output of a filter is a linear weighted combination of the 32 EEG channels. To eliminate noise often present at higher frequencies, so each channel was first band pass filtered in the 0.5–30 Hz range. At each time step, 32-channel EEG data is filtered to provide three filtered outputs. The classifier was trained using this low dimensional filtered time series data. During the BMS procedure, the picture with the highest P300 classifications after all flashes were completed was chosen as the user's choice.

For differentiating between four alternatives, utilizing five flashes per choice, an average classification accuracy of 95% was attained among nine participants. The separate accuracy of each participant is shown in Table 1. With a rate of 4 flashes per second implemented, selecting one of four possibilities takes 5 seconds, resulting in an information transfer rate of 24 bits/min.

Table 1. Participant and accuracy

Participant	Accuracy
1	94.8%
2	93.6%
3	98.2%
4	91.6%
5	93.2%
6	90.6%
7	95.3%
8	98.0%
9	97.6%

4.2. Brain computer interfacing using oscillatory activity

Consider traveling in a virtual environment: left hand, right hand, and foot motor images might be used to move left, right, and forward, respectively. The subject's job in the experiment was to traverse and retrieve coins distributed at random across the surroundings (Figure 4B, panels 2–3). A committee of Fisher's linear discriminant analysis (LDA) classifiers was trained to distinguish between the three kinds of motor imagery. Classification features were calculated from 1 s segments by band pass filtering the EEG signal for numerous frequency bands, squaring, and finding the mean over squared values for each band in each segment. To reduce variability, classification characteristics were based on the logarithm of band power estimations. Each subject's most discriminative frequency bands were discovered individually.

Classification was conducted every 40 ms to allow for real-time engagement. Given the emphasis on motor imagery, data was collected from six EEG sensors positioned over relevant sensorimotor regions. To eliminate EEG signal contamination, techniques for on-line muscle artifact identification and eye movement reduction were also applied. After around 5 h of co-adaptive training over many days, the average 3-class accuracy of the LDA committee classifier reached about 80%, with a false positive rate for motor imagery recognition (by the extra LDA classifier) of around 17%. The BMS was effectively used by the subjects to navigate and locate the coins in the surroundings.

5. Conclusion

The brain machine systems (BMS) system, which transforms brain impulses into messages or commands, uses machine learning. We suggest a novel EEG-based BMS in this study with an emphasis on evoked potential. Nine participants achieved a 95% categorization accuracy on average. An information transfer rate of 24 bits/min is achieved by using a rate of 4 flashes per second, which requires 5 s to choose between four options. Additionally, oscillatory activity used for brain-computer interface was measured. The findings reveal that the average 3-class accuracy of the Linear Discriminant Analysis committee classifier achieved around 80% after approximately 5 h of co-adaptive training over several days, with a false positive rate for motor imagery identification of roughly 17%.

References

[1] A. N. Belkacem, N. Jamil, J. A. Palmer, S. Ouhbi and C. Chen, Brain computer interfaces for improving the quality of life of older adults and elderly patients, Front. Neurosci. 14 (2020) 692.

[2] O. Elemento, C. Leslie, J. Lundin and G. Tourassi, Artificial intelligence in cancer research, diagnosis and therapy, Nat. Rev. Cancer 21 (2021) 747–752.

[3] K. S. Hong, U. Ghafoor and M. J. Khan, Brain–machine interfaces using functional near-infrared spectroscopy: a review, Artif. Life Robot. 25 (2020) 204–218.

[4] N. Jamil, A. N. Belkacem, S. Ouhbi and A, Lakas, Noninvasive electroencephalography equipment for assistive, adaptive, and rehabilitative brain–computer interfaces: a systematic literature review, Sensors 21(14) (2021) 4754.

[5] C. Jeunet, S. Debener, F. Lotte, J. Mattout, R. Scherer and C. Zich, Mind the traps! Design guidelines for rigorous BMS experiments, in: Brain–computer interfaces handbook, CRC Press, 2018, 613–634.

[6] N. E. Khalifa, M. Loey and S. Mirjalili, A comprehensive survey of recent trends in deep learning for digital images augmentation, Artif. Intell. Rev. 55 (2022) 2351–2377.

[7] Y. Li, K. K. Ang and C. Guan, Digital signal processing and machine learning, in: Brain–Computer Interfaces, ed. by B. Graimann et al., Springer-Verlag, Berlin, Heidelberg, 2010, 305–330, https://doi.org/10.1007/978-3-642-02091-9_17.

[8] M. Matheny, S. T. Israni, M. Ahmed and D. Whicher, Artificial intelligence in health care: The hope, the hype, the promise, the peril, Washington, DC, National Academy of Medicine, 2019.

[9] N. Murali and N. Sivakumaran, Artificial intelligence in healthcare–a review, 2018.

[10] W. J. Neumann, M. M. Sorkhabi, M. Benjaber et al., The sensitivity of ECG contamination to surgical implantation site in brain computer interfaces, Brain Stimulation 14(5) (2021) 1301–1306.
[11] A. N. Pisarchik, V. A. Maksimenko and A. E. Hramov, From novel technology to novel applications: Comment on "An integrated brain-machine interface platform with thousands of channels" by Elon Musk and Neuralink, J. Med. Internet Res. 21(10) (2019) e16356.
[12] R. P. N. Rao and R. Scherer, Brain-computer interfacing [in the spotlight], IEEE Signal Processing Magazine 27 (2010) 152–150.
[13] S. Rasheed, A review of the role of machine learning techniques towards brain–computer interface applications, Mach. Learn. Knowl. Extr. 3(4) (2021) 835–862.
[14] A. Roman-Gonzalez, Eeg signal processing for BMS applications, in: Human–Computer Systems Interaction: Backgrounds and Applications 2, Springer, 2012, 571–591.
[15] J. F. D. Saa and M. S. Gutierrez, EEG signal classification using power spectral features and linear discriminant analysis: a brain computer interface application, in: Eighth Latin American and Caribbean Conference for Engineering and Technology, Arequipa, LACCEI, 2010, 1–7.
[16] D. B. Silversmith, R. Abiri, N.F. Hardy, N. Natraj, A. Tu-Chan, E. F. Chang and K. Ganguly, Plug-and-play control of a brain–computer interface through neural map stabilization, Nature Biotechnol. 39(3) (2021) 326–335.

Received on February 7, 2023
Accepted on May 7, 2023

Mohamad Fakhredine and Teodora Ivanova Bakardjieva

Faculty of International Economics and Administration
Varna Free University "Chernorizets Hrabar"
84 Yanko Slavchev Str., Chaika Resort
9007 Varna
BULGARIA

E-mails: `mohamad.fakhredine@vfu.bg`
`bakardjieva@vfu.bg`

ГОДИШНИК НА СОФИЙСКИЯ УНИВЕРСИТЕТ „СВ. КЛИМЕНТ ОХРИДСКИ“
ФАКУЛТЕТ ПО МАТЕМАТИКА И ИНФОРМАТИКА
Том 110

ANNUAL OF SOFIA UNIVERSITY "ST. KLIMENT OHRIDSKI"
FACULTY OF MATHEMATICS AND INFORMATICS
Volume 110

FREQUENCY MODULATED SIGNAL STANDING OUT BY STOCHASTIC RESONANCE EFFECT

OKSANA KHARCHENKO and ZLATINKA KOVACHEVA

In this paper, the phenomenon which is known as stochastic resonance is considered to the signal processing application. Signal standing out with the presence of noise is considered to be one of the basic problems of telecommunication and radioengineering. The stochastic resonance effect is shown to provide significant improvement of some characteristics of the information signal, such as power gain, and noise dispersion at the system output at a certain optimal noise level. In the present article, Minimum Shift Keying signal mixed with Gaussian white noise has been studied using stochastic resonance effect. Noise coefficient, which is one of the quantitative characteristics of noise immunity is calculated and investigated.

Keywords: stochastic resonance, Minimum Shift Keying, signal standing out, noise coefficient, noise dispersion, data engineering

CCS Concepts:

• Applied computing~Physical sciences and engineering

1. Introduction

The main and the most difficult problem of the signal receiving is the problem of noise immunity. The problem of finding the best methods for radio signal receiving with the noise presence is a one of the main goals in radio engineering and telecommunication [12–14].

The researches made in 1980s brought us to paradoxical conclusions. Noise can help amplify the signal. This phenomenon is known as stochastic resonance (SR) [4, 7, 10, 16]. The SR effect presents the response to the weak input signal, in a non-linear system. The system parameters, such as power gain and signal-to-noise ratio (SNR), have their maximum in specific conditions. Nowadays, this effect is

DOI: 10.60063/GSU.FMI.110.95-100

fundamental and has been observed in biology, medicine, psychology, neuroscience, etc. [2, 6, 8, 18].

SR is described by the following equation (see [1]):

$$dy/dt = y(t) - y^3(t) + x(t) + n(t),$$

where $x(t) + n(t)$ is an input process being additive mixture of the useful signal and the white normal noise, and $y(t)$ is the output signal of the non-linear system.

This equation is an Abel equation of first order and it does not have an analytical solution [9].

2. Shift Keying signal standing out from the additive mixture with noise

Currently, telecommunication systems process digitally modulated radio signals. We can cite, as an example, the Decree of the Government of Bulgaria on the widespread use of digital television [11].

Digital modulation is the process of converting digital symbols into signals that are compatible with the characteristics of the channel. Pulses of a given shape modulate the sine wave. Types of modulation depend on the parameter of the radio signal, which varies in proportion to the envelope. Since the digital signal is a sequence of rectangular pulses, and the parameter of the carrier oscillation at a certain point in time changes by a jump, therefore, modulation is actually a manipulation [17].

Currently, the GSM standard uses Minimum Shift Keying (MSK). The primary objective of spectrally efficient modulation techniques is to maximize bandwidth efficiency. Therefore, MSK is widely used in digital telecommunication systems [5, 15].

The modulation of the MSK is as follows: The original data stream consists of bipolar pulses. This pulse stream is divided into an in-phase stream $d_L(t)$ (even bits) and a quadrature stream $d_Q(t)$ (odd bits). The MSK waveform can be expressed as [15]:

$$s(t) = d_L(t) \cos \frac{\pi t}{2T} \cos 2\pi f_0 t + d_Q(t) \sin \frac{\pi t}{2T} \sin 2\pi f_0 t, \tag{2.1}$$

where T is the width pulse and f_0 is the carried frequency.

Let us determine the MSK signal for the pulsed data stream: −1; 1; 1; −1; −1; 1; −1; 1. Figure 1 presents equation (2.1) for this signal. Figure 1a and c illustrate the sinusoidal weighting of the L- and Q-channel pulses, respectively. Figure 1b and d show the modulation of the orthogonal components $\cos \omega_0 t$ and $\sin \omega_0 t$, respectively. Figure 1e illustrates the summation of the orthogonal components from Figure 1b and d.

By adding white Gaussian noise to this signal, we apply the stochastic resonance effect to stand out the useful signal. The result of this signal processing is shown in Figure 2.

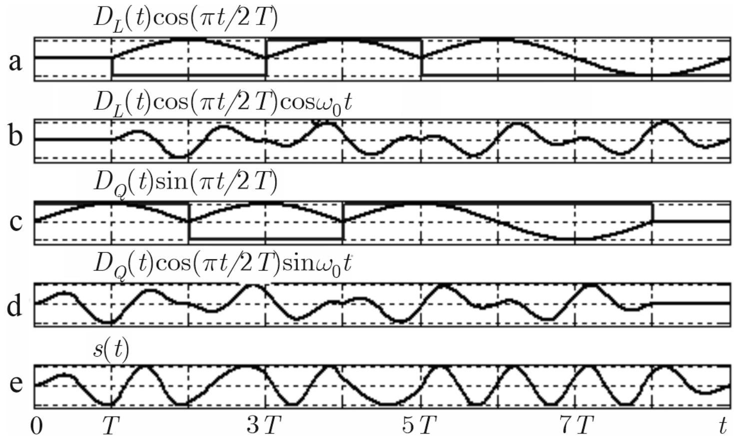

Figure 1. MSK waveform

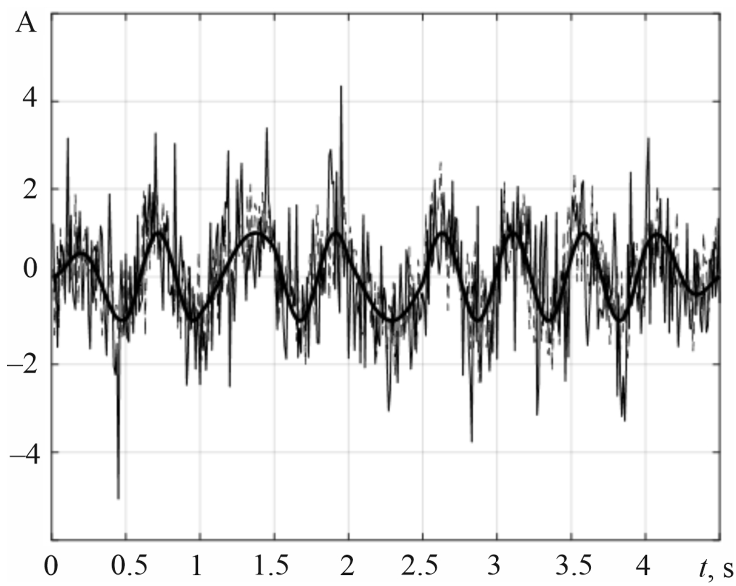

Figure 2. Standing out of the signal from additive signal-noise mixture

It can be seen from Figure 2, that it is possible to significantly reduce the noise component of the oscillations by SR effect. In this case, the input noise dispersion is 1 p.u., the output noise dispersion is 0.4 p.u. The signal is amplified by the SR effect and the power gain is $k = 1.3$.

In Figure 2, input MSK signal is shown by the thick line, signal-noise mixture is illustrated by the thin line, and output signal by the dotted line.

Thus, the SR effect provides significant noise suppression and amplification of the useful signal in communication systems with MSK modulation. Figure 3, which

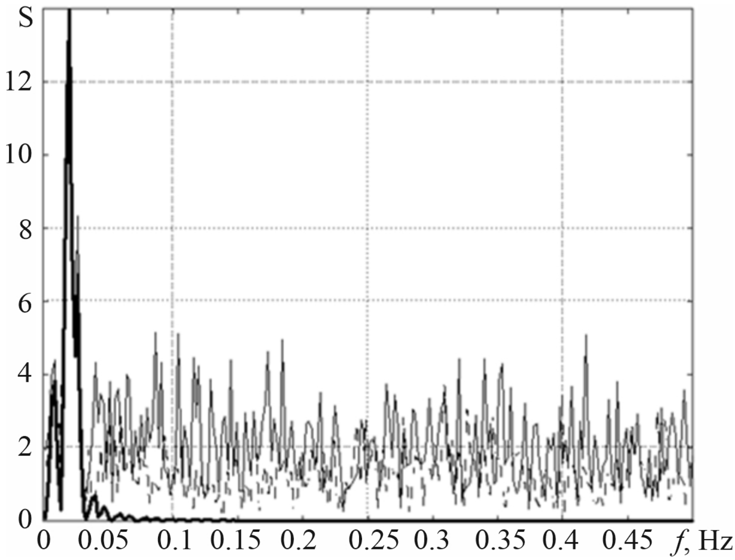

Figure 3. Amplitude spectra (input MSK signal – thick line, signal-noise mixture – thin line, output signal – dotted line)

shows the spectra of these signals, also confirms the effective stand out of the MSK signal.

One of the quantitative characteristics of noise immunity is the noise coefficient. It is defined as a ratio of input signal-to-noise ratio ($\mathrm{SNR}_{\mathrm{input}}$) to output signal-to-noise ratio ($\mathrm{SNR}_{\mathrm{output}}$) [3].

$$L = \mathrm{SNR}_{\mathrm{input}}/\mathrm{SNR}_{\mathrm{output}}. \tag{2.2}$$

We can rewrite this formula in the following form:

$$L = \frac{\mathrm{SNR}_{\mathrm{input}}}{\mathrm{SNR}_{\mathrm{output}}} = \frac{1}{K_p K_N},$$

where:

$K_p = \dfrac{P_{s_{\mathrm{output}}}}{P_{s_{\mathrm{input}}}}$ is the power gain;

$P_{s_{\mathrm{input}}}$ is the input power;

$P_{s_{\mathrm{output}}}$ is the output power;

$K_N = \dfrac{D_{\mathrm{input}}}{D_{\mathrm{output}}}$ is the noise reduction coefficient;

D_{input} is the input noise dispersion;

D_{output} is the output noise dispersion.

From the result, it is seen, that the decrease of the noise coefficient occurs either as a result of an increase in power gain due to an increase in the power of the output signal, or as a result of an increase in the noise reduction coefficient due to a decrease in the output noise level, or as a result of both factors simultaneously. In fact, all these cases present the SR effect, which consists in the transition of noise energy into the energy of a useful signal.

In Figure 2 and 3, $L \approx 0.31$, and the signal is amplified. Therefore, the power gain is greater than 1, the noise reduction factor is also greater than 1 due to the decrease of the output noise level.

3. Conclusion

Stochastic resonance is just beginning to be applied in telecommunication and radioengineering. In this work, the SR effect is used for the first time for standing out of digital signals of the telecommunication systems against the background noise.

The SR effect is investigated, which makes it possible to stand out a weak (compared to noise) signal from an additive mixture of signal and white Gaussian noise.

We showed that SR effect stands out the MSK signal, and provides a significant reduction of the noise component of the signal. As a result, there is an input noise dispersion equals 1 p.u., and the output noise dispersion equals 0.4 p.u. Simultaneously, the signal is amplified by the SR effect and the power gain $k = 1.3$, noise coefficient $L \approx 0.31$.

It is noted, that the decrease of the noise coefficient occurs either as a result of the increase of the power gain, due to the increase of the output signal power, or as a result of the increase of the noise reduction coefficient, due to the decrease of the output noise level.

Acknowledgements

This work was supported by a grant from The Simons Foundation International (Grant No. 992227, IMI-BAS).

References

[1] V. Anishchenko, A. Neiman, F. Moss and L. Schimansky-Geier, Stochastic resonance: noise-enhanced order, Uspekhi Fizicheskikh Nauk, Russian Academy of Sciences 42(1) (1999) 7–34.

[2] A. Bruno and S. Graziani, Stochastic resonance. Theory and application, Kluwer Academic Publishers, 2000.

[3] V. Chumakov, V. Pososhenko, O. Kharchenko and V. Basetsky, Reception and processing of signals (Training manual), KNURE, Kharkiv, 2006, 296 pp (in Ukrainian).

[4] L. Gammaitoni, P. Hänggi, P. Jung and F. Marchesoni, Stochastic resonance: a remarkable idea that changed our perception of noise, Eur. Phys. J. 69(1) (2009) 1–3.

[5] A. D. Hancen, Personal communication networks. Practical implementation, Artech House, Boston–London, 1995.

[6] P. Hänggi, Stochastic resonance in biology, How noise can enhance detection of weak signals and help improve biological information processing, Chemphyschem 12(3) (2002) 90–101.

[7] P. Hänggi and P. Jung, Stochastic resonance, Rev. Mod. Phys. 70(1) (1998) 224–283.
[8] M. Moses Jerome and A. Ramakalyan, Brief survey of stochastic resonance and its application to control, in: 3rd Int. Conf. on Advances in Control and Optimization of Dynamical Systems, Kanpur, India, 2014, 313–321.
[9] E. Kamke, Differentialgleichungen, Bd.1, Gewöhnliche Differentialgleichungen, Teubner Verlag, 10th ed., 1983, 696 pp, ISBN-13: 978-3519120179.
[10] Z. Lai and Y. Leng, Weak-signal detection based on the stochastic resonance of bistable Duffing oscillator and its application in incipient fault diagnosis, Mech. Syst. Signal Process. 8 (2016) 60–74.
[11] Project of a National Strategic Document "Digital Transformation of Bulgaria for the period 2020–2030": https://www.mtc.government.bg/bg/category/167/proekt-na-nacionalen-strategicheski-dokument-cifrova-transformaciya-na-blgariya-za-perioda-2020-2030-g (in Bulgarian).
[12] Y. Shirman et al., Radio-electronic systems: Basics of building and theory, Moscow, ZAO Makvis, 828 pp, 1998 (in Russian).
[13] Y. Shirman et al., Radio-electronic systems: Basics of building and theory (2nd revised Edn), Moscow, Radiotekhnika, 2007 (in Russian).
[14] C. Shraddha, M. L. Chayadevi and M. A. Anusuya, Noise cancellation and noise reduction techniques, in: 1st Int. Conf. on Advances in Information Technology (ICAIT), 2019, 119–123.
[15] B. Sklar, Digital communication. Fundamental and application, 2nd Edn, University of California, Los Angeles, 2001.
[16] U. E. Vincent, P. V. E. McClintock, I. A. Khovanov and S. Rajasekar, Vibrational and stochastic resonances in driven nonlinear systems. Part 2, Philos. Trans. Royal Soc. A 379(2198) (2021) 20210003, https://doi.org/10.1098/rsta.2021.0003.
[17] Yu. I. Voloshchuk, Signals and processes in radio engineering, Textbook for students of higher educational institutions, vol. 2, OO "Company CMIT", Kharkiv, 2003, 102–121, ISBN 966-8530-05-5 (in Ukrainian).
[18] A. Vulpiani, Stochastic resonance in climatic change, Tellus 34(1) (2016) 10–16.

Received on March 26, 2023
Accepted on May 7, 2023

Oksana Kharchenko
Institute of Mathematics and Informatics
Bulgarian Academy of Sciences
Acad. G. Bonchev Str., Bl. 8
1113 Sofia
BULGARIA
E-mail: kharchenko@math.bas.bg

Zlatinka Kovacheva
Institute of Mathematics and Informatics
Bulgarian Academy of Sciences
Acad. G. Bonchev Str., Bl. 8
1113 Sofia
BULGARIA

University of Mining and Geology
Prof. Boyan Kamenov Str.
1700 Sofia
BULGARIA

E-mail: zkovacheva@math.bas.bg

ГОДИШНИК НА СОФИЙСКИЯ УНИВЕРСИТЕТ „СВ. КЛИМЕНТ ОХРИДСКИ“
ФАКУЛТЕТ ПО МАТЕМАТИКА И ИНФОРМАТИКА
Том 110

ANNUAL OF SOFIA UNIVERSITY “ST. KLIMENT OHRIDSKI”
FACULTY OF MATHEMATICS AND INFORMATICS
Volume 110

REAL-TIME DATA INTEGRATION IN INFORMATION SYSTEMS USING STREAM PROCESSING FOR MEDICAL DATA

MARTIN KOSTOV and KALINKA KALOYANOVA

Real-time data processing in medical information systems is becoming harder with the increase in data volume. Stream processing is a popular approach for real-time data processing, which can process large volumes of data including medical in a scalable manner. In some cases, medical data may not be available in real time because of privacy and security concerns. In this paper, we will explore the use of stream processing with static medical data using streaming platforms, Kafka, and Apache Spark. We will demonstrate how these platforms can be used to work with static data in streams and discuss the benefits and limitations of the approach. We also present a case study to illustrate the effectiveness and performance.

Keywords: stream processing, performance comparison, Apache Kafka, Apache Spark, data analysis, medical data

CCS Concepts:

• Information systems~Data management systems~Database management system engines~Stream management

1. Introduction

Real-time data processing is very important for medical information systems (MIS), enabling healthcare providers to access and analyze medical data effectively and fast. Traditional batch processing approaches with relational databases are often inadequate for handling large volumes of medical data, especially in real time. Stream processing has emerged as an alternative approach for real-time data integration, which involves processing and analyzing data in real time as it is generated or ingested.

However, in some cases, medical data may not be available in real-time or may not be suitable for streaming due to privacy and security concerns. In such cases, an

DOI: 10.60063/GSU.FMI.110.101-110

alternative approach can be used – processing static data. Stream processing with static medical data involves processing and analyzing medical data that has been stored in a predefined format – a database, XML files, etc.

This paper examines the use of stream processing with static medical data using stream processing with Apache Kafka [13] and Apache Spark [2]. We discuss the challenges of batch processing in MIS, the advantages of streaming, the main components of stream processing architecture, and some of the characteristics of medical data. We also provide a case study to demonstrate the effectiveness of this approach.

2. Real-time data integration in medical information systems

Medical data is generated continuously. In order to monitor patients adequately, doctors obtain their medical records – diagnoses, medical histories, lab results, images, etc., from different systems. Even when these systems are integrated [3] processing all patient data from different MIS in real-time could be a challenge.

The batch-processing approaches with relational databases are in use most of the time. However, it involves processing data in large batches, which can cause delays in medical data analysis, making it unsuitable for real-time integration in MIS.

Stream processing emerged as a popular approach for real-time data integration in MIS. It enables doctors to monitor patient health in real-time and respond quickly. Stream processing can also provide health trends, enabling healthcare providers to identify patterns and improve patient outcomes, something that is not possible with traditional treatment without hospitalization.

3. Overview of streaming solutions

Stream processing architecture typically consists of several components, including data sources, stream processing frameworks, and output destinations for our tests [1]. We will use Kafka for both input and output. Data sources for static data can include archived medical records or data repositories. Stream processing frameworks, such as Apache Kafka and Apache Spark [4], provide the necessary tools for processing and analyzing static data. Outputs can include dashboards, alerts, and other visualizations that enable doctors to make real-time decisions based on the processed data.

The key benefit of stream processing with static medical data is its ability to handle large volumes of data efficiently with a small overhead. Medical data can be complex and can be from multiple sources, making it difficult to ensure that the data is accurate and consistent. Stream processing with static data enables doctors to monitor this data in real time. With stream processing it is possible to provide insights into patient health trends, enabling doctors to identify patterns and diagnose better.

There are also challenges associated with stream processing static medical data. One of the main challenges is ensuring the quality and consistency of the data. Despite these challenges, stream processing with static medical data can provide significant benefits, including faster response times, improved patient outcomes, and increased efficiency. In the next section, we will explore the characteristics of medical data and the challenges of processing medical data in real time, both with static data and with true streaming data.

4. Stream processing for medical data

Processing medical data in real-time can be difficult due to its specific characteristics - huge volume, different formats, lack of interoperability, etc. But stream processing framework can be utilized to overcome some of these obstacles to enable real-time processing of medical data. This approach can significantly enhance the speed and efficiency of data processing in MIS.

4.1. Medical data distinction

Working with medical data is a complex and diverse field that encompasses a wide range of data sources and data formats. Sometimes this data is organized in predetermined schemas – electronic health records (EHRs), patient summaries, etc. But a huge amount of this data – medical images, sonic representation, physician notes, patient-generated data, and social media is still not organized in predefined schemas.

Several characteristics of medical data make it challenging to process this data in real time. The generated volume is increasing at an unprecedented rate due to advances in medical technology and the adoption of EHRs. The velocity is also increasing, with real-time monitoring of patient data becoming more common [12]. Finally, medical data is characterized by variety, with data generated by a variety of sources, including medical devices, social media, and wearables.

The quality of the data is one of the main challenges to achieve the full processing potential [6]. A basic problem of medical data processing still is the need for standardization of data presentation [8].

4.2. Challenges of processing medical data in real-time

While stream processing can be an effective approach for the real-time processing of medical data, there are also challenges associated with processing static data. One of the biggest challenges is the volume of data, which can be vast and complex, particularly when dealing with large-scale datasets [11]. Processing static data also requires specialized hardware and software infrastructure to handle the volume, complexity, and heterogeneity of the data.

Ensuring data accuracy and consistency is another challenge associated with processing static data. It can be complex and heterogeneous, with data quality issues such as missing data, incomplete data, and data inconsistency leading to inaccurate results and incorrect diagnoses. This makes it important to establish data validation processes and ensure that data is cleaned and normalized before it is processed.

Data privacy and security are critical concerns when processing medical data, particularly when dealing with static data that may be stored for long periods. Healthcare organizations must implement strict security measures to protect sensitive patient information and comply with local regulatory requirements.

Moreover, interoperability between different MIS can be a challenge when working with data [10]. Data may be stored in different formats and structures, making it difficult to exchange and integrate data between different systems [14]. To enable interoperability, a standardized data schema is needed that can accommodate the diverse sources of medical data [9].

To enable real-time processing of medical data, it is also important to have a standardized data schema that can accommodate diverse sources of medical data. A standardized data schema can also enable interoperability between different systems.

5. Problem definition

With more and more data it is becoming challenging for healthcare institutions and researchers to process and analyze all of it in an efficient manner. To address these challenges, there has been growing interest in using stream processing frameworks such as Apache Spark with distributed messaging systems like Apache Kafka to handle the medical data processing in real-time. However, there are still many open questions about how best to design and implement such systems, including issues related to data quality, scalability, fault tolerance, and security. In this paper, we explore the use of stream processing medical data processing and identify key challenges and best practices for designing and implementing such systems.

5.1. Use case dataset for the experiment

In this study, a dataset of 220,000 medical records in XML format is used to simulate a stream of data. The files present daily information about admitted and discharged patients, coming from a hospital – patient demographics, medical history, diagnosis codes, treatment information, etc. The medical records are grouped into files per day for around a period of two years and file size is between 0.5 MB and 3.5 MB depending on the number of patients for the current day.

Since we do not have real-time data, we will use this dataset to simulate a stream that is as close as possible to real-time processing. What we will do is read this dataset, transform it, and publish data to stream. We will measure the mean time it takes to do the reading from the XML file, the transformation, and the publishing to stream. We will also measure how many records we can process in one minute.

5.2. Overview of the experiment environment

The experiment described in the paper uses a setup consisting of a CPU Ryzen 5950 with 4×32 GB DDR4-3200 memory modules under the Ubuntu 22.04 operating system.

We use Apache Kafka – a distributed open-source streaming system where servers and clients communicate via a high-performance TCP network protocol.

```
version: '3.4'
services:
  zookeeper:
    image: confluentinc/cp-zookeeper:7.3.2
    hostname: zookeeper
    environment:
      SERVICE_NAME: zookeeper
      ZOOKEEPER_CLIENT_PORT: 2181
      ZOOKEEPER_TICK_TIME: 2000
    ports:
      - "2181:2181"

  kafka:
    image: confluentinc/cp-kafka:7.3.2
    hostname: kafka
    depends_on:
      - zookeeper
    environment:
      KAFKA_ZOOKEEPER_CONNECT: "zookeeper:2181"
      KAFKA_ADVERTISED_LISTENERS: "PLAINTEXT://kafka:9093,INTERNAL://localhost:9092"
      KAFKA_LISTENERS: "PLAINTEXT://:9093,INTERNAL://:9092"
      KAFKA_LISTENER_SECURITY_PROTOCOL_MAP: "PLAINTEXT:PLAINTEXT,INTERNAL:PLAINTEXT"
      KAFKA_INTER_BROKER_LISTENER_NAME: "PLAINTEXT"
      KAFKA_OFFSETS_TOPIC_REPLICATION_FACTOR: 1
      KAFKA_LOG_SEGMENT_MS: 300000
      KAFKA_LOG_SEGMENT_BYTES: 10485760
      KAFKA_PRODUCER_MAX_REQUEST_SIZE: 2147483648 # 2GB
      KAFKA_CONSUMER_MAX_PARTITION_FETCH_BYTES: 2147483648
      KAFKA_AUTO_CREATE_TOPICS: 1
      KAFKA_NUM_PARTITIONS: 5
      KAFKA_MESSAGE_MAX_BYTES: 1000000000
    ports:
      - "9092:9092"
      - "9101:9101"
```

Figure 1. Apache Kafka deployment configuration

```yaml
version: '3.7'

services:
  spark:
    image: apache/spark:v3.3.2
    container_name: spark
    ports:
      - "8080:8080"
      - "7077:7077"
  postgres:
    image: postgres:15.2
    container_name: postgres-db
    environment:
      POSTGRES_USER: user
      POSTGRES_PASSWORD: password
      POSTGRES_DB: medical_data
    ports:
      - "5432:5432"
```

Figure 2. Apache Spark and PostgreSQL deployment configuration

There are two main operations that clients can perform: publish – write to the stream and consume – read from the stream.

For the purpose of our test, we will use also PostgreSQL – a widespread, open-source object-relational database system, and Apache Spark – an open-source, distributed processing system used for big data workloads.

The two applications – Kafka and Postgre will be evaluated in terms of performance and efficiency when we are working with streams or based on the size of the dataset.

As a tool that defines and runs multiple docker containers, Docker Compose is used. Figure 1 and Figure 2 present the Docker Compose YAML files configuration used for the deployment of Kafka, PostgreSQL, and Spark.

The test environment includes the latest versions of all mentioned tools – Docker v4.18.0, Kafka v7.3.0, Postgres v15.2, and Spark v3.3.2.

5.3. Experiment results and discussion

Overall, the medical dataset used for the experiment contains 220762 records stored in 494 XML files [7]. Every file consists of medical records of daily admitted or discharged patients in a hospital.

We measure how much time it takes to process and publish such a record to stream and how many records we can publish per second. We also tested different batching strategies for the publish to the stream and we see that the publishing strategy is very important.

In the case of batching per file or batching all patient records per day, the results were slowest. If the batching is time-based, the results are better with more time between the batches. The best results were with 75 ms, bigger time between batches was not possible in the existing environment, because the memory of the machine was not enough.

The results of our study demonstrate that Kafka is more suitable for applications that require high throughput, such as real-time monitoring of patient records. By using stream processing with Kafka, healthcare organizations can process medical data in real time, enabling faster and more accurate diagnoses which can improve patient outcomes. Both can be used for Streaming Analytics (SA) and Complex Event Recognition (CER) [5].

Kafka's improved throughput is likely due to its efficient batch-processing capabilities, which allow it to process larger amounts of medical data at once. However, this comes at the cost of higher latency, as it takes longer to process larger batches of medical data. Moreover, if SA or CER is required the throughput will be reduced.

When we process the dataset file by file and publish it to Kafka the throughput is not optimal as seen in the first row in Table 1. On the other hand, if we use time-based batching the results are better – the second row.

Table 1. Publishing the medical records to Kafka stream measurements

Publish batching	Publish per second	Publish mean time (in ms)
Per XML file	4136	0.24
Time-based batching (75 ms)	13 872	0.072

Based on our results we present the following recommendations:

- When working with **small to medium** datasets **without streaming** medical data – Kafka or Spark are not needed as they could lead to more downsides than benefits.
- When working with **small to medium** datasets **with streaming** of medical data – Kafka can be used as an extension or relational database management system such as PostgreSQL.
- When working with **large datasets** – a system like Spark is required as it has the capabilities to process large datasets. Since Kafka does not provide data integrity reliable storage is still required.

In Table 2 we have summarized some of the main advantages and disadvantages of the different platforms we tested.

Healthcare organizations should carefully consider their specific needs and requirements when choosing a stream processing platform and evaluate the performance of different platforms using their medical data and specific use cases.

Table 2. Aggregation source advantages and disadvantages

	Spark C# with **Postgres** source	Spark C# with **Kafka** source
Processing time for aggregation	Faster	Slower due to additional I/O
Scalability	Limited by PostgreSQL capabilities	Easy to add more instances
Data integrity	Guaranteed with Postgres	Potential for data loss
Use case data size	Low to medium data volumes	High data volumes with scalability

6. Conclusions

In this study, we discuss the importance of choosing the right stream processing tools for specific cases of data usage, as well as the need for continued development and improvement in real-time data integration for medical information systems. We tested different options for fast processing medical data. We compared two different data sources – Kafka and PostgreSQL. Overall, stream processing tools such as Kafka and Spark are important for handling large amounts of data or when real-time processing with low latency is important. When real-time processing is not needed or the dataset is not very big, then those tools may not be needed, and the traditional approach is a good decision.

Future research in this area could focus on the development of new stream processing tools that are specifically designed for handling medical data in real time, as well as the integration of machine learning and artificial intelligence techniques for real-time data analysis and decision-making. Additionally, further testing and evaluation of existing stream processing tools for medical data could provide valuable insights into their performance and limitations.

Acknowledgements

This work is supported by the Fund for Scientific Research at Sofia University "St. Kliment Ohridski" under Grant 80-10-8/11.04.2023 and the project BG05M2OP 001-1.001-0004 (UNITe) funded by Operational Program Science and Education for Smart Growth co-funded by European Regional Development Fund.

References

[1] A. Akanbi and M. Masinde, A distributed stream processing middleware framework for real-time analysis of heterogeneous data on big data platform: Case of environmental monitoring, Sensors 20(11) (2020) 3166, https://doi.org/10.3390/s20113166.

[2] M. Armbrust, R. S. Xin, C. Lian et al., Spark SQL: Relational data processing in Spark, in: SIGMOD'15: Proc. 2015 ACM SIGMOD Int. Conf. on Management of Data (2015) 1383–1394, https://doi.org/10.1145/2723372.2742797.

[3] A. Bocevska, S. Savoska, I. Jolevski, N. Blazheska-Tabakovska and B. Ristevski, Implementation of innovative e-Health services and digital healthcare ecosystem – Cross4all Project summary, in: Proc. Information Systems & Grid Technologies (ISGT2022), CEUR Workshop Proceedings, 285–301, http://ceur-ws.org/Vol-3191/paper26.pdf.

[4] F. Hassan, M. E. Shaheen and R. Sahal, Real-time healthcare monitoring system using online machine learning and Spark streaming, Int. J. Adv. Comput. Sci. Appl. 11(9) (2020) 650–658, https://doi.org/10.14569/IJACSA.2020.0110977.

[5] S. Langhi, R. Tommasini and E. D. Valle, Extending Kafka streams for complex event recognition, in: 2020 IEEE Int. Conf. on Big Data (2020) 2190–2197, https://doi.org/10.1109/BigData50022.2020.9378217.

[6] K. Kaloyanova, I. Naydenova and Z. Kovacheva, Addressing data quality in healthcare, in: Proc. Information Systems and Grid Technologies (ISGT2021), CEUR Workshop Proceedings, 155–164, http://ceur-ws.org/Vol-2933/paper16.pdf.

[7] M. Kostov, StreamingMedicalData (2023) https://github.com/KostovMartin/StreamingMedicalData.

[8] P. Kovachev, E. Krastev, D. Tcharaktchiev, E. Markov and I. Evg. Ivanov, Conversion of Bulgarian observational data to OMOP common data model: Initial results, in: Proc. Information Systems & Grid Technologies (ISGT'2022), CEUR Workshop Proceedings, 113–125, http://ceur-ws.org/Vol-3191/paper10.pdf.

[9] E. Krastev, S. Abanos and D. Tcharaktchiev, Health data exchange based on archetypes of clinical concepts, in: Proc. Information Systems & Grid Technologies (ISGT2022), CEUR Workshop Proceedings, 98–112, http://ceur-ws.org/Vol-3191/paper09.pdf.

[10] M. Nisheva, H. Georgiev and P. Pavlov, Building a semantic repository for outpatient sheets, in: Proc. Information Systems and Grid Technologies (ISGT 2020), CEUR Workshop Proceedings, 14–29, http://ceur-ws.org//Vol-2656/paper3.pdf.

[11] I. Patias and V. Georgiev, The use of big data in medicine and public health policymaking: Opportunities and challenges, in: Proc. Information Systems & Grid Technologies (ISGT2020), CEUR Workshop Proceedings, 7–13, https://ceur-ws.org/Vol-2656/paper1.pdf.

[12] S. Savoska, A. Bocevska and H. Simonoski, Visualization of sensors' data in time series databases for health purposes, in: Proc. Information Systems & Grid Technologies (ISGT'2022), CEUR Workshop Proceedings, 317–334, http://ceur-ws.org/Vol-3191/paper28.pdf.

[13] M. J. Sax, G. Wang, M. Weidlich and J.-C. Freytag, Streams and tables: Two sides of the same coin, in: BIRTE'18: Proc. Int. Workshop on Real-Time Business Intelligence and Analytics (2018) 1–10, https://doi.org/10.1145/3242153.3242155.

[14] S. Velikov, K. Merdzhanov, N. Leventi and T. Kundurdzhiev, Application of openEHR platform for data exchange in ophthalmology, in: Proc. Information Systems & Grid Technologies (ISGT'2022), CEUR Workshop Proceedings, 126–134, http://ceur-ws.org/Vol-3191/paper11.pdf.

Received on March 31, 2023
Accepted on May 7, 2023

Martin Kostov

Faculty of Mathematics and Informatics
Sofia University "St. Kliment Ohridski"
5 James Bourchier Blvd.
1164 Sofia
BULGARIA

E-mail: martinkk@fmi.uni-sofia.bg

Kalinka Kaloyanova

Faculty of Mathematics and Informatics
Sofia University "St. Kliment Ohridski"
5 James Bourchier Blvd.
1164 Sofia
BULGARIA

Institute of Mathematics and Informatics
Bulgarian Academy of Sciences
Acad. G. Bonchev Str., Bl. 8
1113 Sofia
BULGARIA

E-mail: kkaloyanova@fmi.uni-sofia.bg

ГОДИШНИК НА СОФИЙСКИЯ УНИВЕРСИТЕТ „СВ. КЛИМЕНТ ОХРИДСКИ“
ФАКУЛТЕТ ПО МАТЕМАТИКА И ИНФОРМАТИКА
Том 110

ANNUAL OF SOFIA UNIVERSITY “ST. KLIMENT OHRIDSKI”
FACULTY OF MATHEMATICS AND INFORMATICS
Volume 110

PREPROCESSING TECHNIQUES FOR BRAIN MRI SCANS: A COMPARATIVE ANALYSIS FOR RADIOGENOMICS APPLICATIONS

MARIYA MITEVA and MARIA NISHEVA-PAVLOVA

In this study, we aim to investigate the use of preprocessing techniques on brain magnetic resonance imaging (MRI) scans for the prediction of Methylguanine-DNA methyltransferase methylation (MGMT) status in glioma patients. MGMT methylation is a biomarker that has been linked to treatment response and prognosis in glioma. We review several studies that have applied preprocessing techniques to brain MRI scans, along with molecular genetic information, for this purpose. The preprocessing techniques include but are not limited to image registration, normalization, brain extraction, and tumor segmentation. We compare the effectiveness of the techniques used in these studies and evaluate the performance of each technique in terms of accuracy, computational efficiency and other parameters. Our goal is to identify the most effective preprocessing techniques for radiogenomics applications and to determine the potential of these techniques for improving the accuracy of predictions in brain MRI scans by combining different types of data. The results of this study have the potential to serve as a basis for the development of more accurate and efficient imaging-based diagnostic tools for glioma patients, and to improve the understanding of the relationship between imaging and genomics in glioma.

Keywords: radiogenomics, glioblastoma, MGMT promoter, MRI scans, medical imaging, machine learning, deep learning

CCS Concepts:

- Computing methodologies∼Artificial intelligence∼Computer vision∼Computer vision representations∼Image representations;
- Computing methodologies∼Artificial intelligence∼Computer vision∼Computer vision problems∼Image segmentation;
- Computing methodologies∼Machine learning∼Machine learning approaches∼Neural networks

DOI: 10.60063/GSU.FMI.110.111-125

1. Introduction

Magnetic resonance imaging (MRI) is a widely used modality for non-invasive imaging of the brain and other organs. It provides high-resolution, multi-dimensional images of the tissue anatomy and function, which can be used for various clinical and research purposes. One of the emerging areas of research is radiogenomics, which aims to integrate imaging data with other types of genomic and clinical data to improve the diagnosis, prognosis, and treatment of diseases [18]. For example, the prediction of Methylguanine-DNA methyltransferase methylation (MGMT) status in glioblastoma, a type of brain cancer, has been shown to be a promising marker for personalized medicine and targeted therapy [20].

MGMT is a Deoxyribonucleic acid (DNA) repair enzyme that removes alkyl groups from DNA, and it has been shown to be involved in the resistance of cancer cells to chemotherapy [11]. MGMT methylation status has been shown to be associated with the response to chemotherapy and the prognosis of several types of cancer, including glioblastoma, the most common and aggressive primary brain cancer [11]. There are several methods available to evaluate MGMT promoter methylation, including methylation-specific polymerase chain reaction (MSP), multiplex ligation-dependent probe amplification (MLPA), pyrosequencing (PSQ), quantitative Real-Time PCR, and immunohistochemistry (IHC) to assess protein expression [19].

Medical imaging, such as brain MRI, can provide valuable information about the brain tissue and its structural and functional characteristics, which can be useful for predicting MGMT methylation status and other genomic and clinical phenotypes [16]. However, medical images are often affected by noise, artifact, and intensity variations, which can degrade the quality and accuracy of the images, and hinder the performance of the predictive models [15]. Preprocessing techniques aim to improve the quality and accuracy of the medical images, by removing the noise, artifact, and intensity variations, and by enhancing the tissue characteristics [17].

In addition to improving the quality and accuracy of images, preprocessing techniques also play an important role in the success of Computer-aided detection and diagnosis (CAD) schemes. CAD schemes have become an increasingly important tool in medical imaging to help clinicians read images more efficiently and make diagnoses more accurately and objectively. The use of CAD schemes in medical imaging is not a new concept, with early CAD schemes being developed in the 1970s. However, the development of CAD schemes has accelerated in recent years, particularly since the 1990s [8], due to the integration of more advanced machine learning methods. Conventional CAD schemes typically involve three steps: target segmentation, feature computation, and disease classification. Target segmentation is the process of identifying and isolating the region of interest (ROI) in the image, such as a tumor or lesion. Feature computation involves quantifying the characteristics of the ROI in terms of size, morphology, margin geometry, texture, and so on. Finally, disease classification involves using a classification model, such as linear discrimination analysis (LDA), to identify the malignancy of the ROI.

Deep learning-based models have become increasingly popular in recent years and have shown promising results in CAD schemes. These models involve a hierarchical architecture that can learn important features hidden in the raw image in a self-taught manner, eliminating the need for manual feature development [14]. In deep learning-based models, a neural network is trained to identify and amplify important features related to the specific task, while filtering out irrelevant features. This process is done in a progressive manner, with the neural network gradually recognizing and learning more complex features as the number of layers increases [4].

A number of studies have compared the performance of different preprocessing techniques for medical images, but most of them have focused on specific types of images and applications, such as CT, PET, or ultrasound [4,9]. It can be concluded that there is a lack of comprehensive and comparative studies on the performance of preprocessing techniques for brain MRI scans in the context of radiogenomics applications, such as predicting MGMT methylation status [1,5,6,12]. This paper aims to fill this gap by providing a comparative analysis of various preprocessing techniques for brain MRI scans, using a radiogenomics dataset and assesses the impact of the preprocessing techniques on the results obtained from deep and machine learning algorithms.

2. Contemporary research methodologies in predicting MGMT methylation status

2.1. Findings and study

Several studies have employed various preprocessing techniques on brain magnetic resonance imaging (MRI) scans to enhance the predictive capabilities of determining methylation status of the O6-methylguanine-DNA methyltransferase gene in brain tumors. The selection of the six studies involved in the comparative analysis of contemporary research methodologies in predicting MGMT methylation status is based on several criteria related to their relevance, scientific rigor, and diversity of methodologies employed. One key criterion for inclusion is the scientific quality and significance of the studies. The selected studies are likely to have undergone rigorous peer-review processes and have been published in high-impact journals, ensuring the reliability and validity of their findings. Another important criterion is the diversity of methodologies employed in the studies. The selected studies are likely to have employed a range of different imaging techniques and data preprocessing methods to enhance the predictive capabilities of determining MGMT methylation status. By including studies with different methodologies, the comparative analysis can provide a more comprehensive overview of the current state of research in the field, identify the strengths and limitations of different approaches, and highlight potential areas for further improvement. The techniques discussed aim to improve the quality and information content of the imaging data for more accurate and efficient assessment of MGMT methylation status. For example, the authors in [2] employed a series of image preprocessing steps to prepare the imaging data for analysis. The primary goal of the preprocessing was to align and standardize the imaging data across different modalities for each patient. First, the authors used the FMRIB Linear Image Registration Tool (FLIRT) for coregistration of the imaging data across different modalities. The FLIRT algorithm is based on linear affine transformation and uses a mutual-information cost function. The reference volume for coregistration was the highest resolution sequence, most commonly the postcontrast T1-weighted acquisition. On average, coregistration of a single volume took approximately 1 minute.

In this study, MR imaging data and corresponding molecular genetic information were retrospectively obtained from the Cancer Imaging Archives and the Cancer Genome Atlas for patients with low- or high-grade gliomas. Only patients with full preoperative MR imaging, including T2, FLAIR, and T1-weighted pre- and postcontrast acquisitions, were included in the analysis. The molecular information for each patient was obtained, including IDH1 status, 1p/19q codeletion, and MGMT promoter methylation.

Subsequently, each input image was independently normalized using z-score normalization. This step is commonly used to standardize the data, and to ensure that the mean of the data is zero and the standard deviation is one. The authors then employed a custom in-house fully automated whole-brain extraction tool, based on 3D convolutional neural network, to remove extracranial structures. This step aimed to improve computational efficiency and focus the analysis on relevant regions.

Finally, the authors used a fully automated brain tumor segmentation tool to identify lesion margins. The algorithm used in this step was the top-performing tool as evaluated in the international 2016 Multimodal Brain Tumor Segmentation Challenge. It is based on a serial fully convolutional neural network architecture with residual connections, and it performs whole-tumor segmentation in approximately 1 second. These segmentations were used to generate cropped slice-by-slice images of the tumor on all modalities, which were subsequently resized to a $32 \times 32 \times 4$ input.

It is worth noting that the use of convolutional neural network (CNN) based algorithm was developed and tested for the purpose of predicting genetic mutations and methylation status in glioma patients. The algorithm was found to have high accuracy in predicting IDH1 mutation (mean 94%), 1p/19q codeletion (mean 92%), and MGMT promoter methylation (mean 83%), as well as good performance in terms of the area under the curve for these predictions. The CNN was trained for 25,000 iterations before convergence, and the entire imaging workflow takes approximately 5.12 s per patient, which includes time for detection, preprocessing, and classification. This approach demonstrated promising results and may have potential clinical applications.

In a subsequent experiment, discussed in [3], S. Chen et al. obtained approval from the local Institutional Review Board and recruited 111 patients diagnosed with WHO grade 2-4 glioma who had undergone surgical resection and received plain and enhanced scans from the Affiliated Drum Tower Hospital of Nanjing University Medical School between 2018 and 2020. The patients had not received any prior treatment such as radiotherapy, chemotherapy or antitumor drugs before surgery and those with incomplete or poor-quality images were excluded. The data were divided into a training group and a validation group with a ratio of 8:2, and MRI images acquired using 3.0 T MRI scanners and different protocols were used in the analysis. The image brightness variations due to scanning process were eliminated and deviation field correction was performed on all images before the analysis.

In addition to the previously described methodology, the authors also employed a method for extracting radiomics features from the tumor edema and tumor core area of four sequences, including T1WI, T2WI, T1CE, and ADC. They used a combination of manual segmentation by two radiologists, and an open-source platform called PyRadiomics to extract a total of 688 features from each patient. These features were then normalized by the Z-score and used as input in the deep learning model. The image sequences were registered to the same physical space in order to match the same patient ROIs across each sequence. After completing the preprocessing steps, the study subjects were randomly assigned to either the training or testing group in an 80:20 ratio.

In this study, a ResNet deep learning model based on radiomics features was used to predict the MGMT promoter methylation status of gliomas. The study found that among single MRI modalities, the T1CE model based on the Region of Interest (ROI) of the tumor core achieved the highest AUC value of 0.84. When multiple MRI modalities were combined, the T1CE model combined with the ADC model based on the ROI of the tumor core achieved the highest AUC value of 0.90. The final model, which was a combination of T1CE and ADC modalities, based on the ROI of the tumor core, showed the best performance among all the models, with the highest accuracy of 0.91 and AUC of 0.90. Ten features were found to be the most important radiomics features for the prediction. The study suggests that the combination of T1CE and ADC MRI modalities could be superior to other single or multiple MRI sequences in the prediction of MGMT promoter methylation and that a deep learning model based on radiomics features could help in identifying molecular biomarkers from routine medical images, and therefore facilitate treatment planning.

In another work [10], the authors used brain MRI scans from patients diagnosed with glioblastoma multiforme (GBM) in order to analyze the relationship between methylation status and imaging characteristics. GBM is a highly malignant brain tumor with a poor prognosis, and identifying biomarkers that can inform diagnosis and treatment strategies is a crucial area of research.

The MRI scans used in the study were obtained from the Cancer Imaging Archive (TCIA) and consisted of 5,235 scans from 262 patients. Each scan was a 3-dimensional reconstruction of the brain and were provided in a DICOM format, which is a non-proprietary data interchange protocol, digital image format, and file structure for biomedical images and related information. These scans were preprocessed to remove noise by looking at the distribution of Hounsfield Units in the pixels and only retain the slices that contain the tumor. Additionally, all images were resized to 128x128 dimensions for consistency.

The study also utilized methylation data from the Cancer Genome Atlas (TCGA) for 423 unique patients, which were preprocessed to extract methylation sites that are located in minimal promoter and enhancer regions. These regions are known to have a high level of methylation activity and affect MGMT expression, a DNA repair protein that is commonly inactivated in GBM. Specifically, the study focused on three methylation sites cg02941816, cg12434587, and cg12981137 which had been used in previous studies on MGMT methylation and GBM. A patient was considered to have positive methylation status if any of the three sites had a methylation beta value of at least 0.2.

The study also employed data augmentation techniques to increase the size of the dataset and prevent overfitting in the convolutional recurrent neural network (CRNN) used in the analysis. The data augmentation involved applying image rotation and reversing the MRI scans, so that the methylation status and location of the tumor was preserved. The images were rotated every 4° from -90° to $+90^\circ$, and were flipped so that the MRI scans were represented from superior to inferior and vice versa. This resulted in a 90-fold increase in the number of MRI scans available for training the CRNN, which is expected to boost the performance and robustness of the network.

The study found that the CRNN obtained modest patient-level accuracy of 0.67 on the validation set and 0.62 on the test set. It was also better in performance when compared to the random forest model. The CRNN also provided a generalizable platform for visualizing the different filters and layers of deep learning architectures for brain MRI scans.

In a subsequent research, P. Korfiatis et al. [13] emphasize the importance of reducing manual steps required by computer-aided diagnosis systems in order to facilitate the translation of such systems into clinical practice. The proposed system uses image normalization and bias corrections as the only preprocessing steps, which are fully automated and computationally efficient, taking less than 2 minutes on a typical desktop computer. The system also focuses on a three-class problem rather than a binary approach, enabling the algorithm to operate without the need for tumor segmentation step, thus reducing the complexity of the process.

This study aims to investigate the relationship between methylation status and imaging characteristics in patients with newly diagnosed GBM using MRI scans. The study was approved as minimal risk by the institution's Internal Review Board and included 155 presurgery MRI examinations from patients treated at Mayo Clinic between 2007 and 2015. The inclusion criteria were age ≥ 18 years and preoperative MR scans with T2 and T1 weighted post-contrast images performed at Mayo Clinic with known MGMT methylation status. The images were anonymized and the image processing pipelines were managed with MIRMAID. The study found 66 patients had methylated and 89 patients had unmethylated tumors. For the methylated group, 53 scans were performed on a 1.5T scanner and 13 were performed on a 3T scanner, while for the unmethylated group, 76 scans were performed on a 1.5T scanner and 13 were performed on a 3T scanner. For the purpose of this study, only T2 images were used and N4 was used for bias field correction to eliminate the low-frequency and smooth signal that corrupts MRI images and potentially affect image analysis steps.

The authors evaluated the ability of three different residual deep neural network (ResNet) architectures to predict methylation status of the O6-methylguanine methyltransferase gene using magnetic resonance imaging without the need for a distinct tumor segmentation step. The study found that the ResNet50 architecture (50 layers) was the best performing model, achieving an accuracy of 94.90% for the test set. This performance was statistically significantly better than both ResNet18 (18 layers) and ResNet34 (34 layers) architectures. This study proposes a method that eliminates the need for extensive preprocessing and demonstrates that deep neural architectures can be used to predict molecular biomarkers from routine medical images.

The authors of [21] use multiparametric MRI images of brain gliomas from the TCIA and genomic information from the TCGA and TCIA. The final dataset of 247 subjects included 163 methylated cases and 84 unmethylated cases. Tumor masks for 179 subjects were obtained through previous expert segmentation and for the remaining 68 subjects, they were generated by trained 3D-IDH network and reviewed by two neuroradiologists. The preprocessing steps included: 1) Affine coregistration using Advanced Normalization Tools, 2) skull stripping using Brain Extraction Tool (BET), 3) removal of radiofrequency inhomogeneity using N4 Bias Field Correction, and 4) normalization of intensity to zero-mean and unit variance. The entire preprocessing took about 5 min per dataset.

In this study, transfer learning was applied to predict the MGMT promoter status utilizing a previously trained 3D-IDH network. The decoder part of the network was fine-tuned for voxel-wise dual-class segmentation of the whole tumor, where one represents methylated and two represent unmethylated MGMT promoter types. The authors used a dataset of multiparametric MR images of patients with brain gliomas obtained from the TCIA database, in combination with genomic information obtained from both the TCGA and TCIA databases. The data preprocessing steps applied to the images included: 1) affine coregistration to the SRI24 T2 template using the Advanced Normalization Tools software package, 2) skull stripping using the BET from the Oxford Centre for Functional MRI of the Brain Software Library (FSL), 3) removing radiofrequency inhomogeneity using N4 Bias Field Correction, and 4) normalizing intensity to zero-mean and unit variance. In order to assess the network's performance, the authors implemented a 3-fold cross-validation strategy. The dataset of 247 subjects was randomly shuffled and distributed into three groups, and then the three groups alternated among training, in-training validation, and held-out testing groups. The network's performance was reported only on the hold-out testing group for each fold because it is never seen by the network during the training.

Finally, a 3D-IDH network was trained and fine-tuned for determination of MGMT promoter status using transfer learning method. A three-fold cross-validation approach was implemented on a dataset of 247 subjects with MRI images and known MGMT promoter status. The network achieved a mean testing accuracy of 94.73% across the three folds with a range of 93.98–95.12% (SD 0.66%). Additional evaluation metrics such as sensitivity, specificity, positive predictive value, negative predictive value and AUC were also computed with high performance, ranging from 91.66% to 96.31% (SD 2.06%). The network showed an average Dice score of 0.82

(SD 0.008) for tumor segmentation. The network misclassified 13 cases among the total of 247 subjects.

The authors in [7] aim to build a model for predicting the methylation status of the O6-Methylguanine-DNA-methyltransferase promoter in glioblastoma multiform patients, using a novel radiomics-based machine learning (ML) approach. The following steps were used to build the model:

1. Data source and collection: The pre-processed and segmented multimodal magnetic resonance imaging features from the Cancer Genome Atlas – GBM24 collections were downloaded from the Cancer Imaging Archive public database. Only data entries with MRI modalities such as T1-weighted pre-contrast (T1), T1-weighted post-contrast (T1-Gd), T2, and T2-FLAIR (fluid-attenuated inversion recovery) were selected, resulting in 53 GBM patients included in the study. The 704 radiomics features obtained were classified into seven categories: first-order statistical features, volumetric features, textural features, histogram-based features, morphological features, spatial features, and glioma diffusion properties.

2. A two-stage radiomics feature selection and machine learning classification approach was used to predict the MGMT methylation status in glioblastoma and low-grade glioma (LGG) patients using medical imaging data. Three machine learning models (Random Forest (RF), XGBoost, and Support Vector Machine (SVM)) were incorporated into a genetic algorithm (GA) algorithm for feature selection.

The GA-RF model was found to have the best performance with a sensitivity of 0.894, specificity of 0.966, and accuracy of 0.925 in the GBM dataset. The GA-RF feature set outperformed other feature selection methods with an AUC of 0.93 in identifying MGMT methylation status from radiomics features. In the LGG dataset, the GA-RF model outperformed other models with an accuracy of 0.750, sensitivity of 0.78, and specificity of 0.62. The results indicate the potential of applying the extracted radiomics features for the prediction of MGMT methylation status in both high- and low-grade gliomas.

2.2. Conclusions

In these studies, various preprocessing techniques are used to enhance the predictive capabilities of determining methylation status of the MGMT gene in brain tumors. The preprocessing of magnetic resonance imaging scans is a crucial step in improving the quality and increasing the information content of the imaging data. This is done to obtain a more precise assessment of brain tumors. The commonly employed preprocessing techniques include image registration, normalization, extraction of the entire brain, and segmentation of the tumors. These techniques are executed using software tools such as the FMRIB Linear Image Registration Tool and custom algorithms that utilize convolutional neural networks. Another method for preprocessing involves the extraction of radiomics features from various MRI sequences, which can be performed using either manual segmentation or open-source platforms like PyRadiomics. These features are then used as inputs for deep and machine learning models, such as ResNet, to predict the methylation status of the O6-methylguanine-DNA methyltransferase gene in brain tumors. The results from these studies indicate a promising level of performance and have potential applications in the clinical setting.

3. Discussion

3.1. Discoveries and analysis

In this section we present a comparison, which summarizes the key characteristics of studies investigating the prediction of MGMT methylation status from brain MRI scans.

The studies mentioned above have used deep or machine learning models to predict various glioma subtypes based on medical imaging data and genetic information. The research in Table 1 provides information about the dataset, preprocessing techniques, model architecture, and evaluation metrics used in the studies, arranged in a clear and organized manner. The studies vary in their specific focus, but all aim to use machine learning and in particular deep learning models to predict different glioma subtypes with high accuracy.

Based on the comparison provided in the table, the authors in [13] reported an accuracy of 94.90% which is the highest among the other studies in the table. Furthermore, they used a dataset of 155 pre-surgery MRI examinations from patients treated at Mayo Clinic between 2007 and 2015, which is a relatively small dataset compared to other studies such as [2] and [10] that used datasets from the Cancer Imaging Archive. The study used only T2 images from the MRI exams, and applied normalization and bias correction as the only preprocessing steps, which were fully automated and computationally efficient. Additionally, the study used ResNet50 architecture, which is a 50-layer deep convolutional neural network, and trained it to classify the images as either methylated or unmethylated tumors. The study found that the ResNet50 architecture performed the best out of all the ResNet architectures tested, achieving an accuracy of 94.90% on the test set. One possible reason for the high performance of the ResNet50 architecture in this study is the small size of the dataset used. With a small dataset, it is possible that the model is able to achieve high performance by overfitting to the training data. Moreover, the study not used any additional data except MRI images, that makes the results only based on the ability of the model to extract the relevant information from the images.

The work of the researchers in [7] is noteworthy for its use of a unique combination of imaging and genomic data which allows for a more comprehensive analysis and prediction of MGMT methylation status. They used a combination of multimodal MRI scans (T1, T1-Gd, T2, and T2-FLAIR) and radiomics features (extracted from

Table 1

Comparative analysis of MRI preprocessing and modeling techniques in radiogenomics

Authors. Chang et al. [2]
Year. 2018
Dataset. MR imaging data and corresponding molecular genetic information retrospectively obtained from the Cancer Imaging Archives and the Cancer Genome Atlas for patients with low- or high-grade gliomas.
Preprocessing. FLIRT for coregistration of the imaging data across different modalities. The FLIRT algorithm is based on linear affine transformation and uses a mutual-information cost function.
Model. 3D Convolutional Neural Network (CNN)
Accuracy. High accuracy in predicting IDH1 mutation (mean 94%), 1p/19q codeletion (mean 92%), and MGMT promoter methylation (mean 83%)
Validation. IDH1 mutation (mean, 94%; range between cross validations, 90– 96%), 1p/19q codeletion (mean, 92%; range, 88–95%), and MGMT promoter methylation (mean, 83%; range, 76– 88%) on 5-fold cross-validation

---

Authors. Han & Kamdar [10]
Year. 2018
Dataset. MRI scans were obtained from the TCIA and consisted of 5,235 scans from 262 patients; also methylation data from the TCGA for 423 unique patients are used, which were preprocessed to extract methylation sites.
Preprocessing. Scans were preprocessed to remove noise by looking at the distribution of Hounsfield Units in the pixels and only retain the slices that contain the tumor; all images were resized to 128×128 dimensions for consistency.
Model. Convolutional Recurrent Neural Network (CRNN)
Accuracy. Patient-level accuracy of 0.67 on the validation set and 0.62 on the test set
Validation. Examined predictions on the test set

Authors. Korfiatis et al. [13]
Year. 2017
Dataset. 155 presurgery MRI examinations from patients treated at Mayo Clinic between 2007 and 2015; 66 patients had methylated and 89 patients had unmethylated tumors.
Preprocessing. Normalization and bias corrections as the only preprocessing steps, which are fully automated and computationally efficient.
Model. ResNet50 architecture (50 layers) is the best performing model
Accuracy. 94.90% accuracy on the test set
Validation. k-fold cross validation

Authors. Yogananda et al. [21]
Year. 2021
Dataset. MRI images of brain gliomas from the TCIA and genomic information from the TCGA and TCIA
Preprocessing. 1)Affine coregistration using Advanced Normalization Tools, 2) Skull stripping using Brain Extraction Tool, 3) Removal of radiofrequency inhomogeneity using N4 Bias Field Correction, 4) normalizing intensity to zero-mean and unit variance
Model. 3D-IDH network
Accuracy. Mean testing accuracy of 94.73% across the 3 folds with a range of 93.98–95.12% (SD 0.66%), AUC ranging from 91.66% to 96.31% (SD 2.06%)
Validation. 3-fold cross validation

Authors. Chen et al. [4]
Year. 2022
Dataset. 111 patients diagnosed with WHO grade 2–4 glioma who had undergone surgical resection and received plain and enhanced scans from the Affiliated Drum Tower Hospital of Nanjing University Medical School between 2018 and 2020.
Preprocessing. A combination of manual segmentation, and an open-source platform called PyRadiomics to extract features from each patient; then Z-score normalization is applied. The image sequences were registered to the same physical space in order to match the same patient ROIs across each sequence.

Model. ResNet
Accuracy. The final model, which was a combination of T1CE and ADC modalities, with the highest accuracy of 0.91 and AUC of 0.90
Validation. 5-fold cross validation

Authors. Do et al. [7]
Year. 2022
Dataset. Text53 GBM patient data from the Cancer Imaging Archive database with MRI modalities such as T1, T1-Gd, T2, and T2-FLAIR. 704 radiomics features were extracted and classified into 7 categories: statistical, volumetric, textural, histogram-based, morphological, spatial, and diffusion properties.
Preprocessing. Already pre-processed and segmented multimodal magnetic resonance imaging (MRI)
Model. 3 ML models (Random Forest (RF), XGBoost, and Support Vector Machine (SVM)) were incorporated into a GA algorithm for feature selection.
Accuracy. The GA-RF model was found to have the best performance with a sensitivity of 0.894, specificity of 0.966, and accuracy of 0.925 in the GBM.
Validation. 5-fold cross validation

---

the scans and classified into seven categories). The authors utilized three machine learning models (Random Forest, XGBoost, and Support Vector Machine) and incorporated them into a genetic algorithm for feature selection. The results showed that the GA-RF model had the best performance with a sensitivity of 0.894, specificity of 0.966, and accuracy of 0.925 on the GBM dataset, based on 5-fold cross validation.

The results of this study demonstrate the potential for using machine learning techniques in the analysis of brain MRI scans and radiogenomic data. The combination of imaging and genomic information has the potential to improve the accuracy of patient outcomes prediction, which can inform treatment decisions and improve patient outcomes. Future studies in this direction can focus on further refining the genetic algorithm for feature selection and incorporating additional machine learning models. Additionally, larger datasets can be used to validate the results and explore the potential for using these techniques in a clinical setting. The study provides a significant contribution to the field of radiogenomics and highlights the importance of combining imaging and genomic data to improve patient outcomes.

3.2. Conclusions

The main findings from the comparison tables are closely related to the topic of this paper, which is preprocessing brain MRI scans for predicting MGMT methylation status. The findings suggest that deep and machine learning-based techniques and multi-modal approaches may be more effective for this task, and that larger and more diverse datasets may be more useful for training and evaluating such models. By considering these findings, researchers can make more informed decisions about the most appropriate preprocessing techniques and datasets to use for their specific research questions and goals. Additionally, these findings may also be useful for clinicians and healthcare professionals who are interested in using imaging and other data to predict and manage the treatment of brain tumors in their patients.

4. Incorporating expert knowledge in medical imaging analysis

Incorporating expert knowledge is a crucial aspect for the development of deep and machine learning algorithms for radiogenomics applications using brain MRI scans. The ability to explain and understand the decision-making processes of these algorithms is essential for their clinical implementation and acceptance. Expert knowledge, specifically domain knowledge, can aid in the optimization of imaging data and improve the performance of predictive models.

Embedding expert knowledge into the preprocessing and analysis of brain MRI scans can address the unique challenges present in medical images, such as high inter-class similarity and limited labeled data. For example, incorporating anatomical information can improve the registration of multi-modal imaging data, while incorporating radiomic features can enhance the representativeness of the imaging data. Additionally, incorporating text reports accompanying images can provide additional clinical information for the decision-making process [4].

Furthermore, incorporating expert knowledge into the training and validation process can also improve the interpretability of the models. This can be achieved through the use of methods such as feature importance analysis and decision tree visualization. These methods allow for the identification of the most important features used by the model in its decision-making process and can provide insight into how the model is using the expert knowledge. This can help researchers and clinicians understand how the model is making its predictions, which can ultimately lead to more trust in the model's predictions. Additionally, the use of expert knowledge can also lead to the development of more robust models that are able to generalize better to unseen data.

It is also worth mentioning that expert knowledge does not need to be only from a radiologist, other experts from different fields such as computer vision, medical physics, or medical informatics can also bring valuable insights to optimize the algorithms. For example, computer vision experts can help in preprocessing the images, medical physics experts can help in understanding the underlying physics of the images and medical informatics experts can help in understanding the clinical context of the images. Collaboration between different experts can lead to a more comprehensive approach to radiogenomics.

In conclusion, incorporating expert knowledge in medical imaging analysis is crucial for the development of deep and machine learning algorithms for radiogenomics applications using brain MRI scans. Integrating domain expertise into the preprocessing, analysis, and validation of the data can lead to more accurate and interpretable models that are better suited for clinical implementation. Collaboration between experts from different fields can also bring valuable insights that can optimize the algorithms further.

5. Conclusions

In conclusion, the use of preprocessing techniques for brain MRI scans has been shown to be a useful tool for radiogenomics applications, particularly in the prediction of MGMT methylation status. A comparative analysis of the studies listed in the table, such as [2, 3, 7, 10, 13, 21], reveal that each study used different preprocessing techniques, models and achieved different levels of accuracy. The results discussed in [7] is particularly noteworthy for its use of a unique combination of imaging and genomic data, which allows for a more comprehensive analysis and prediction of MGMT methylation status, and the high accuracy achieved by the GA-RF model with a sensitivity of 0.894, specificity of 0.966, and accuracy of 0.925 based on 5-fold cross validation of the GBM dataset. This study demonstrates the potential for combining multiple data sources to improve predictions in medical imaging.

It has been demonstrated that the integration of various types of data can provide a more comprehensive understanding of the underlying biology of brain tumors and potentially enhance the diagnostic and therapeutic decision-making process. Therefore, it is important for future research to focus on investigating the potential of incorporating multiple data modalities in radiogenomics applications, in addition to the development of advanced preprocessing techniques that can optimize the quality and information content of imaging data for improved prediction accuracy. Furthermore, with the advancement of deep learning techniques, there is a growing potential to integrate these models into preprocessing techniques, which can lead to more accurate and efficient predictions, and also enable to explain the decision-making process. It is important to consider the use of deep and machine learning models in future research in order to fully exploit the potential of radiogenomics applications in the diagnosis and treatment of brain tumors.

References

[1] D. Abler, V. Andrearczyk, V. Oreiller et al., Comparison of MR preprocessing strategies and sequences for radiomics-based MGMT prediction, in: Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 7th Int. Workshop, BrainLes 2021, Held in Conjunction with MICCAI 2021, Virtual Event, September 27, 2021, Revised Selected Papers, Part II, Springer-Verlag, Berlin, Heidelberg, 367–380. https://doi.org/10.1007/978-3-031-09002-8_33.

[2] P. Chang, J. Grinband, B. Weinberg et al., Deep-learning convolutional neural networks accurately classify genetic mutations in gliomas, Am. J. Neuroradiol. 39(7) (2018) 1201–1207.

[3] S. Chen, Y. Xu, M. Ye et al., Predicting MGMT promoter methylation in diffuse gliomas using deep learning with radiomics, J. Clin. Med. 11 (2022) 3445, https://doi.org/10.3390/jcm11123445.

[4] X. Chen, X. Wang, K. Zhang et al., Recent advances and clinical applications of deep learning in medical image analysis, Med. Image Anal. 79 (2022) 102444, https://doi.org/10.1016/j.media.2022.102444.

[5] X. Chen, M. Zeng, Y. Tong et al., Automatic prediction of MGMT status in glioblastoma via deep learning-based MR image analysis, Biomed. Res. Int. 2020 (2020) 9258649, https://doi.org/10.1155/2020/9258649; PMID: 33029531; PMCID: PMC7530505.

[6] K. Dang, T. Vo, L. Ngo and H. Ha, A deep learning framework integrating MRI image preprocessing methods for brain tumor segmentation and classification, IBRO Neurosci. Rep. 13 (2022) 523–532, https://doi.org/10.1016/j.ibneur.2022.10.014; PMID: 36590099; PMCID: PMC9795279.F.

[7] D. Do, M. Yang, L. Lam et al., Improving MGMT methylation status prediction of glioblastoma through optimizing radiomics features using genetic algorithm-based machine learning approach, Sci. Rep. 12 (2022) 13412, https://doi.org/10.1038/s41598-022-17707-w.

[8] K. Doi, H. MacMahon, S. Katsuragawa et al., Computer-aided diagnosis in radiology: potential and pitfalls, Eur. J. Radiol. 31 (1999) 97–109.

[9] N. Goel, A. Yadav and B. M. Singh, Medical image processing: A review, in: 2016 2nd Int. Innovative Applications of Computational Intelligence on Power, Energy and Controls with their Impact on Humanity (CIPECH), Ghaziabad, India, 2016, 57–62, https://doi.org/10.1109/CIPECH.2016.7918737.

[10] L. Han and M. Kamdar, MRI to MGMT: predicting methylation status in glioblastoma patients using convolutional recurrent neural networks, in: Pacific Symp. Biocomputing 2018: Proc Pacific Symp., World Scientific, 331–342, 2018.

[11] M. Hegi, A. Diserens, T. Gorlia et al., MGMT gene silencing and benefit from temozolomide in glioblastoma, N. Engl. J. Med. 352(10) (2005) 997–1003, https://doi.org/10.1056/NEJMoa043331; PMID: 15758010.

[12] K. Jae, L. Pyo, K. Jae and K. Gi, Comparison of pre-processed brain tumor MR images using deep learning detection algorithms, J. Multimed. Inf. Syst. 8(2) (2021) 79–84, https://doi.org/10.33851/JMIS.2021.8.2.79.

[13] P. Korfiatis, T. Kline, D. Lachance et al., Residual deep convolutional neural network predicts MGMT methylation status, J. Digit. Imaging 30(5)(2017) 622–628.

[14] Y. LeCun, Y. Bengio and G. Hinton, Deep learning, Nature 521 (2015) 436–444.

[15] S. Masoudi, S. Harmon, S. Mehralivand et al., Quick guide on radiology image preprocessing for deep learning applications in prostate cancer research, J. Med. Imaging (Bellingham) 8(1) (2021) 010901, https://doi.org/10.1117/1.JMI.8.1.010901; PMID: 33426151; PMCID: PMC7790158.

[16] B. Menze, A. Jakab, S. Bauer et al., The multimodal brain tumor image segmentation benchmark (BRATS), IEEE Trans. Med. Imaging 34(10) (2015) 1993–2024, https://doi.org/10.1109/TMI.2014.2377694; PMID: 25494501; PMCID: PMC4833122.

[17] M. Salvi, U. Acharya, F. Molinari and K. Meiburger, The impact of pre- and post-image processing techniques on deep learning frameworks: A comprehensive review for digital pathology image analysis, Comput. Biol. Med. 128 (2021) 104129, https://doi.org/10.1016/j.compbiomed.2020.104129.

[18] L. Shui, H. Ren, X. Yang et al., The era of radiogenomics in precision medicine: An emerging approach to support diagnosis, treatment decisions, and prognostication in oncology, Front. Oncol. 10 (2021) 570465, https://doi.org/10.3389/fonc.2020.570465; PMID: 33575207; PMCID: PMC7870863.

[19] L. Wang, Z. Li, C. Liu et al., Comparative assessment of three methods to analyze MGMT methylation status in a series of 350 gliomas and gangliogliomas, Pathol. Res. Pract. 213(12) (2017) 1489–1493, https://doi.org/10.1016/j.prp.2017.10.007; PMID: 29103769.

[20] M. Weller, R. Stupp, G. Reifenberger et al., MGMT promoter methylation in malignant gliomas: ready for personalized medicine?, Nat. Rev. Neurol. 6(1) (2010) 39–51.

[21] C. Yogananda, B. Shah, S. Nalawade et al., MRI based deep-learning method for determining glioma MGMT promoter methylation status, Am. J. Neuroradiol., 42(5) (2021) 845–852, https://doi.org/10.3174/ajnr.A7029, http://www.ajnr.org/content/early/2021/03/04/ajnr.A7029.

Received on March 17, 2023
Accepted on May 7, 2023

Mariya Miteva

Faculty of Mathematics and Informatics
Sofia University "St. Kliment Ohridski"
5 James Bourchier Blvd.
1164 Sofia
BULGARIA

E-mail: mmiteva@fmi.uni-sofia.bg

Maria Nisheva–Pavlova

Faculty of Mathematics and Informatics
Sofia University "St. Kliment Ohridski"
5 James Bourchier Blvd.
1164 Sofia
BULGARIA

Institute of Mathematics and Informatics
Bulgarian Academy of Sciences
Acad. G. Bonchev Str., Bl. 8
1113 Sofia
BULGARIA

E-mail: marian@fmi.uni-sofia.bg

ГОДИШНИК НА СОФИЙСКИЯ УНИВЕРСИТЕТ „СВ. КЛИМЕНТ ОХРИДСКИ“
ФАКУЛТЕТ ПО МАТЕМАТИКА И ИНФОРМАТИКА
Том 110

ANNUAL OF SOFIA UNIVERSITY "ST. KLIMENT OHRIDSKI"
FACULTY OF MATHEMATICS AND INFORMATICS
Volume 110

GAME THEME BASED INSTRUCTIONAL MODULE TO TEACH LOOPS AND CHOICE STATEMENTS IN COMPUTER SCIENCE COURSES

MITKO TRPKOSKI, ANDRIJANA BOCEVSKA, IGOR NEDELKOVSKI and SNEZANA SAVOSKA

With the fast development of computer science and technology, computer games have become one of the integral parts of modern way of living. Research studies have conveyed that educational games are motivating, engaging and provide a reliable learning context. Due to the expansion of the educational transform, game theme-based learning methodology has become one of the current research focuses. There is a need to change the traditional passive method of teaching to an active method of teaching such as game and simulation-based learning. So, we need to provide a better learning environment by increasing the student's motivation towards learning. The game theme-based instructional (GTI) modules prepare the learners to think critically, and the students can adopt new challenges of the relevant knowledge. This paper presents a novel and exciting methodology of teaching algorithms by motivating the students towards learning. We designed and developed The Ball Targeting Game consisting of two GTI modules to teach loops and choice statements. The gaming modules demonstrated in this paper were developed using Vizard, which is a virtual reality toolkit used for developing virtual worlds and immersive applications for visualization and simulation using Python as its scripting language.

Keywords: Game Theme Based Instructional Module, computer science, algorithms

CCS Concepts:

• Software and its engineering~Software creation and management~Software development techniques

1. Introduction

Teaching computer science and coding to students can often be a difficult task and many of the concepts and skills involved can seem complex. This is usually

DOI: 10.60063/GSU.FMI.110.127-138

the case for algorithms, which are an essential part of computer science, but many students can find them confusing if they are not taught effectively. The traditional way of teaching is text-based programming, which is not so inspiring to most of the students. Teaching algorithms using Game theme-based instructional (GTI) module can provide a better understanding of the concept than with a traditional instruction approach and engage students to learn by hands-on experience. Due to the change in the education system, five of the most frequently studied nontraditional methods are:

- Flipped classroom
- Problem-based learning
- Gamification
- Case studies and
- Social media-centered learning.

The authors of [9, 10, 12] defined the flipped classroom as a "pedagogical model in which the typical lecture and homework elements of a course are inverted". The definition of problem-based learning according to [7,20] is "an instructional method in which students learn through facilitated problem-solving." Gamification is described as "the use of game design elements in non-game contexts" [3, 14]. The case study method refers to "an approach that allows researchers to develop and present an in-depth view of a particular situation, event or entity" [19]. The researchers of [5,8] defined social media-centered as "a group of Internet-based applications that build on the ideological and technological foundations of Web 2.0, and that allow the creation and exchange of user-generated content.

Game theme-based learning helps to improve the problem-solving skills of students and produce better learning outcomes [21,23]. GTI modules help the students to have a better understanding of concepts by engaging them more towards learning according to [16, 18, 23]. Studies have shown that the simulation element of games has educational potential regarding both subjective and social perspectives [22, 24].

This paper discusses the design and implementation of the two Game theme-based instructional (GTI) modules named Shipment and Factory for teaching loops and choice statements to computer science students in an introductory programming course. The Ball Targeting Game was developed to teach loops and choice statements. The GTI modules are developed using virtual reality software and provide better understanding of the concept of loops and choice statements. The gaming modules demonstrated in this paper were developed using Vizard, which is a virtual reality toolkit used for developing virtual worlds and immersive applications for visualization and simulation using Python as its scripting language. Vizard supports high-quality 3D sound and multi-user networking and can import other gaming environments by adding them in the script. The main objective of the GTI module is to provide better understanding of the concept of loops and choice statements by using a constructivist approach of learning by doing.

The paper is structured as follows. The next section briefly describes the work done previously. Section 3 discusses the modeling of GTI modules, while in Section 4 the implementation of GTI modules is presented. Finally, we conclude and provide insights for future work in Section 5.

2. Related works

The design, implementation, and evaluation of a game theme-based instructional (GTI) module to teach linked list and binary tree is presented by [17]. GTI modules are designed to invigorate the instructors to teach and motivate the learners to learn the concepts of the linked list and binary trees using a gaming metaphor. The purpose of their paper is to explore the issues concerning the usability and likability of game theme-based instructional module to teach linked list and binary trees. The results of the evaluation of GTI modules show the effectiveness of GTI modules and demonstrates that the GTI module is more usable and likable. Game theme based instructional modules for computer science students that motivate and engage students while contributing to their learning outcomes are developed by [18]. They design game theme based instructional modules to encourage faculty to teach and motivate students to learn the concepts of object-oriented programming using interactive, graphical, game-like examples. Study of [1] focuses on designing and implementing a virtual reality (VR) game-based application (iThinkSmart) to support computational thinking (CT) knowledge. The study followed the design science research methodology to design, implement, and evaluate the first prototype of the VR application. An initial evaluation of the prototype was conducted with 47 computer science students from a Nigerian university who voluntarily participated in an experimental process. Authors of the paper [21] designed and implemented two modules teaching arrays and loops aiming to provide an alternative way of teaching the two data structures to students in introductory Computer Science courses. To accomplish this, they developed the two Virtual reality instructional (VRI) modules within the Vizard development platform using Python and included seven different types of loops for the Duck Game and two functions for the Array game. The purpose of the study presented in [2] is to investigate the use of an educational game to enhance student learning effectiveness. The results show that the combination of gamification and traditional learning methods can enhance students' learning motivation and learning effects. According to their conclusion, with the development of educational technology, educational games are becoming more and more popular, and further research is therefore recommended. The paper [4] surveys gamification tools used to teach computer programming. According to them, many researchers use game mechanics in a story to teach, encourage, and engage students in learning programming concepts. Also, this technique is a useful support for instructors because it helps them to realize more engagement, motivation, collaboration, fun, and effectiveness. Six serious games with various genres for teaching information security courses and evaluate their effectiveness as an efficient teaching tool are presented by [13]. The study also determines which game genre is the most suitable for delivering educational contents. The obtained results proved and confirmed the hypothesis that educational games have a positive impact as a pedagogic tool on the educational process. In their work, [6] present a collaborative game that was developed to assist students in learning algorithms and they explore its learning capabilities. The game aims to assist students in learning constraint satisfaction algorithms and it is based on the map coloring game. They conducted an evaluation study in real classroom conditions and revealed quite promising results which indicate that the game is an effective way to enhance students' motivation, engagement and interest and it also helps students to deeper understand the functionality of constraint satisfaction algorithms.

3. Modeling the Game Theme-based Instructional (GTI)

This section discusses the modeling of the two GTI modules: Shipment and Factory. It gives a brief overview of the Vizard platform used to develop the games as well as the concepts covered in the Ball Targeting Game.

3.1. Design considerations

The modules were built with the constructivist theory in mind, which states that the students build knowledge by experiencing it for themselves in the real world. Constructivism refers to the belief that the learners construct knowledge for themselves via interaction. Constructivists focus more on an understanding of knowledge through experience and less on verifying the concept [15]. The modules were also built using functional and nonfunctional requirements. The functional requirements were taken from the student's perspective while the nonfunctional requirements were taken from the instructor's perspective. The functional requirements include the following:

- User interface must be intuitive,
- The student should be able to understand the instructions,
- Displayed pseudocode and flowchart algorithm should be comprehensible.

The nonfunctional requirements include the following:

- Each module should motivate the student to learn,
- Each module should teach the student about the loops and choice statements,
- Reaction to user input should be immediately rendered,
- Graphics should be appealing to the students,
- The modules should be portable, i.e. the student should be able to play each module on any platform,
- An award system should be featured.

3.2. Vizard framework

Vizard as a virtual reality toolkit is a Python-based integrated development environment (IDE) used to develop virtual reality applications. Three-dimensional models can be built in 3D MAX and then imported into the Vizard environment using its built-in exporter. Models and images imported into Vizard can be positioned into the environment and scaled to fit. Through its libraries, Vizard provides built-in functions that govern interactions between objects and their environments. One can also add shapes, text, buttons, and sliders through those functions. Vizard consists of a Python script editor and a debugger. After a script has been created, it can be run with or without debugging. It can also be published as an executable (.exe) file for use by the general public.

3.3. Loop and choice statement GTI module concept

The Ball Targeting Game covers different types of loops and choice statements. They are listed as follows:

1. **If/Else.** In this type of choice statement, an action is carried out if a condition is met. If the condition is not satisfied, a different action is carried out.
2. **Nested If/Else.** This type of choice statement is executed the same way as an **If/Else** choice statement; the only difference is that this choice statement is an **If/Else** choice statement located within another type of choice statement or loop.
3. **Switch statements.** In this type of statement, if a variable being evaluated meets a certain condition, then an action is executed for that condition. Otherwise, a default action is executed.
4. **While loop.** The actions specified in the body of this loop are repeated as long as a certain condition (loop guard) is satisfied. The loop breaks when the condition is no longer met.

4. Game theme-based instructional (GTI) module implementation

This section details the implementation of the two modules and the user interaction featured within them.

4.1. Game engines

The Ball Targeting Game was implemented using the Python programming language within the Vizard Virtual Reality Toolkit, a Python-based integrated development environment used to develop virtual reality applications. On the developer's side, the Vizard IDE provided the libraries needed for the source code of the two modules, represented as ballTargetingGame.py. After development, an executable file ballTargetingGame.exe was created from the Python file. Once the executable file has been created, it could be used on a standalone computer, without having Vizard installed on it beforehand.

4.2. The Ball Targeting Game

In this section, the entire process of creating a foundational application will be explained in which through a game student will understand virtual reality coding as the basis of a game. The main idea is to create two modules of the game called Shipment and Factory. In both modules, multiple targets are set and essentially multiple functions are repeated. The essence is to have some kind of weapon with which we will need to hit all of the given targets that make some kind of movements without touching them. We need to find the targets which are mapped and placed in certain positions on the scene itself. In order to make it easier to understand and display the functions, the application is divided into several classes. Ready-made models are also used, which are imported into Vizard and thus used during implementation to realize our goal. Following libraries are needed to provide access to the basic Vizard functions:

- import viz – library which is necessary because it provides access to the Vizard library,
- import vizcam – library with which we call the camera that will follow the model,
- import vizact – library that included commonly used applications,
- import vizshape – library with which we will be able to position the models,
- import math – library with which we will be able to provide different values of different constants.

In order to achieve the goal of enhancing students' learning about algorithms and in particular loops and choice statements, a flowchart diagram corresponding to the interaction flow in the game has been created and given in Figure 1. Students need to study the flowchart to obtain an initial understanding on how they can play the game and win. By comprehending the flow of the game based on the algorithm, students can learn how the different types of choice statements and loops behave and how they can be integrated in a computer program. After students have understood the flow and the general idea of the concepts, the game code can be made available to them as a next step. This would provide them the possibility to apply the gained knowledge by modifying the existent loops and choice statements or by interchanging concepts by related ones (e.g., one type of loop with another type of loop) and observing if the flow of the game has changed. We strongly believe that this approach will be especially helpful for computer science students without previous knowledge or programming experience, as in this case they would be given an initial program, which they can modify and further enhance instead of having to

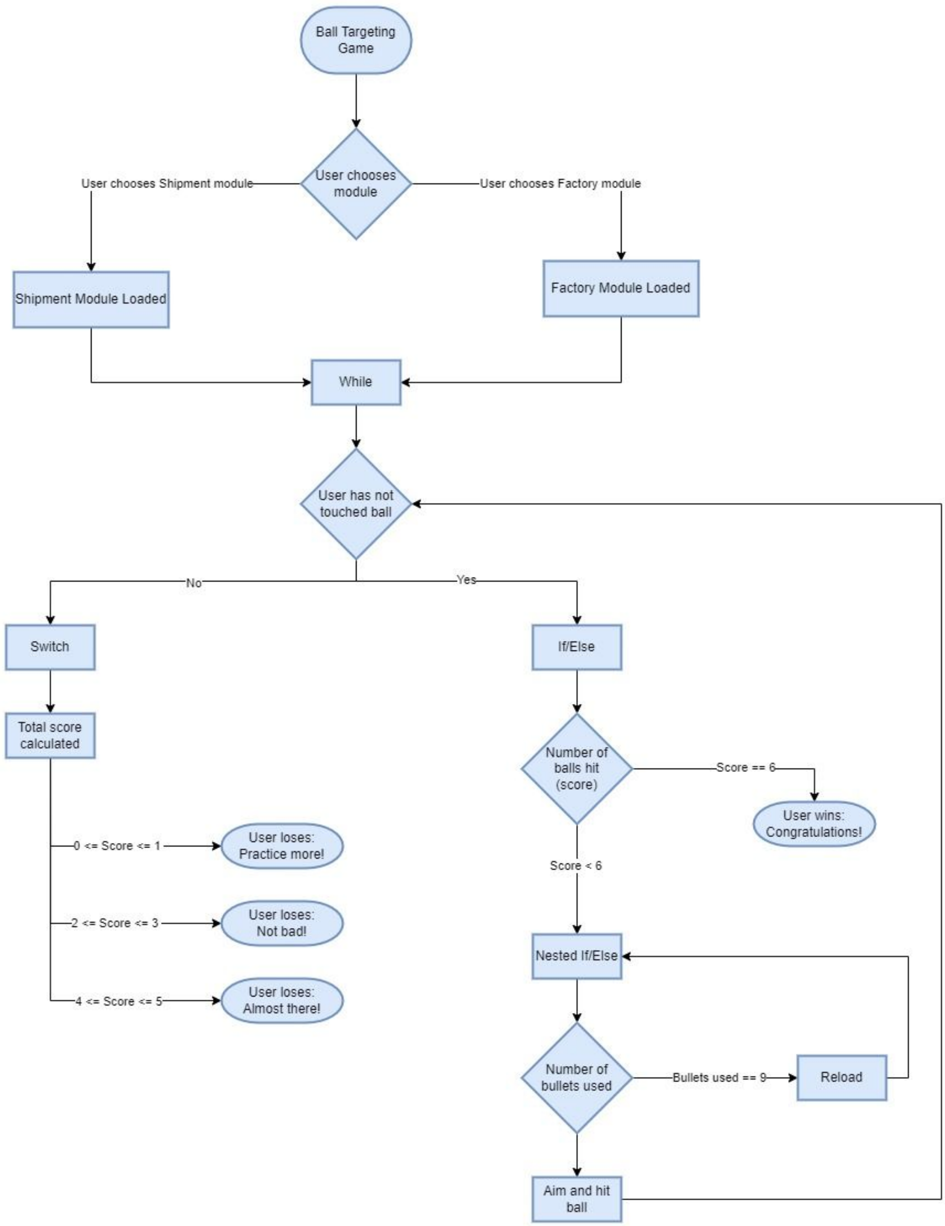

Figure 1. Flowchart representing the different loops and choices offered in the Ball Targeting Game

develop everything by themselves from the very beginning. The students' motivation and the effectiveness of the approach are planned to be evaluated in the future using a questionnaire.

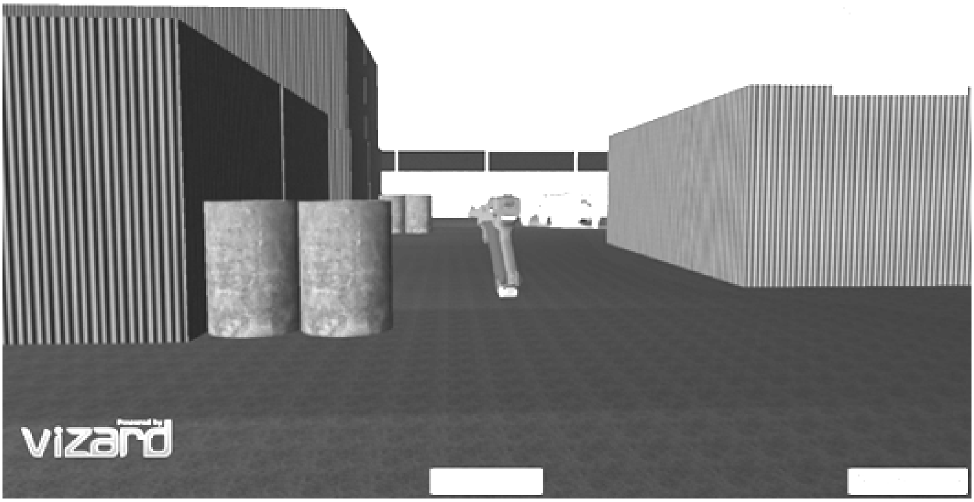

Figure 2. User screen after the Shipment module is loaded

Figure 3. User screen after the Factory module is loaded

The interaction of the user with the game is as follows. At the beginning, a prompt asking the user to choose a module is shown. After the selection of a module, the respective module is loaded and an instruction about how the user can navigate is displayed, as shown in Figure 2 and Figure 3.

In each module there are six balls in total which need to be found and hit by the player. The player can move inside the map as described in the navigation shown in the corner of the screen and can try to hit a ball with a bullet by pressing the Space

Figure 4. Reloading bullets when no more are available

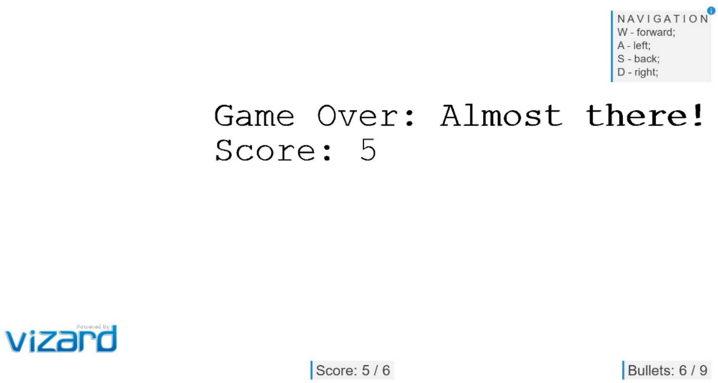

Figure 5. One of the screens shown when the user loses the game

button. The player has initially 9 bullets on disposal, but in case all of them has been used, the Button R can be pressed in order to reload new bullets, as shown in Figure 4.

The choice statements and loops are explained as follows:

- **While.** This loop contains the guard which determines when the game stops. As long as the user has not touched a ball, the game is running and the user can search for the balls, try to hit them and reload the gun if necessary. If a user touches a ball, the game stops and the user has lost the game.

- **Switch.** If the user loses the game, this option determines "how good" the user was in the game. Here, the number of balls hit before losing the game is considered and a message such as “Practice more!”, “Not bad!” or “Almost there!” is displayed (Figure 5).

- **If/Else.** Using this statement it is checked if the user has successfully found and hit all six balls. If the condition is fulfilled, the user wins and a respective message is shown. Otherwise, the game continues.
- **Nested If/Else.** While playing the game, the user occasionally has the option to reload new bullets if the condition that all nine available bullets have been used is met.

5. Conclusion

We designed and implemented two modules teaching loops and choice statements aiming to provide an alternative way of teaching algorithms in introductory Computer Science courses. To accomplish this, we developed the two VRI modules within the Vizard development platform using Python. In both modules, multiple targets are set and essentially multiple functions are repeated. The modules were built with the constructivist theory in mind, which states that the students build knowledge by experiencing it for themselves in the real world. The modules were also built considering functional and nonfunctional requirements. Students can study the provided flowchart algorithm prior to playing the game in order to understand the behaviour of different loops and choice statements and learn how they can be integrated in a computer program. As a further work, GTI modules evaluation using questionnaire to find the motivation of students and the effectiveness of the approach is planned. Modern technologies like automatized attendance tracking systems for tracking students’ attendance and obtaining their feedback can be used [11]. We hope that the results of the evaluation of GTI modules will show that instructional modules can efficiently promote learning by encouraging the students’ participation. Furthermore, we plan to develop more GTI modules to teach other concepts including, but not limited to concepts like linked lists, binary trees and arrays to introduce students to an alternate method of learning.

References

[1] F. J. Agbo, S. S. Oyelere, J. Suhonen and M Tukiainen, Design, development, and evaluation of a virtual reality game-based application to support computational thinking, Educ. Tech. Res. Dev., 71 (2023) 505–537, https://doi.org/10.1007/s11423-022-10161-5.

[2] S. Y. Cheung and K. Y. Ng, Application of the educational game to enhance student learning, Front. Educ. 6 (2021) Art. 623793, 10 pp, https://doi.org/10.3389/feduc.2021.623793.

[3] S. Deterding, D. Dixon, R. Khaled and L. Nacke, From game design elements to gamefulness: Defining “gamification”, in: Proc. Int. Acad. MindTrek Conf.: Envisioning Future Media Environments, Tampere, Finland, 28–30 September 2011, 9–15, https://doi.org/10.1145/2181037.2181040.

[4] R. Elshiekh and L. Butgerit, Using gamification to teach students programming concepts, Open Access Library J., 4 (2017) e3803, https://doi.org/10.4236/oalib.1103803.

[5] J. Gikas and M. M. Grant, Mobile computing devices in higher education: Student perspectives on learning with cellphones, smartphones & social media, Internet High. Educ. 19 (2013) 18–26, https://doi.org/10.1016/j.iheduc.2013.06.002.

[6] F. Grivokostopoulou, I. Perikos and I. Hatzilygeroudis, A collaborative game for learning algorithms, in: Proc. 9th Int. Conf. on Computer Supported Education (CSEDU 2017), vol. 1, 2017, 543–549, https://doi.org/10.5220/0006377405430549.

[7] C. E. Hmelo-Silver, Problem-based learning: What and how do students learn?, Educ. Psychol. Rev. 16 (2004) 235–266, https://doi.org/10.1023/B:EDPR.0000034022.16470.f3.

[8] A. M. Kaplan and M. Haenlein, Users of the world, unite! The challenges and opportunities of Social Media, Bus. Horiz. 53(1) (2010) 59–68, https://doi.org/10.1016/j.bushor.2009.09.003.

[9] A. Karabulut-Ilgu, N. J. Cherrez and C. T. Jahren, A systematic review of research on the flipped learning method in engineering education, Br. J. Educ. Technol. 49 (2018) 398–411, https://doi.org/10.1111/bjet.12548.

[10] S. C. Kong and Y. Song, An experience of personalized learning hub initiative embedding BYOD for reflective engagement in higher education, Comp. Educ. 88 (2015) 227–240, https://doi.org/10.1016/j.compedu.2015.06.003.

[11] Z. Kotevski, N. B. Tabakovska, A. Bocevska and T. Dimovski, On the technologies and systems for student attendance tracking, Int. J. Inf. Technol. Comput. Sci. 10(10) (2018) 44–52, https://doi.org/10.5815/ijitcs.2018.10.06.

[12] M. J. Lage, G. J. Platt and M. Treglia, Inverting the classroom: A gateway to creating an inclusive learning environment, J. Econ. Educ. 31(1) (2000) 30–43, https://doi.org/10.1080/00220480009596759.

[13] M. Mostafa and O. S. Faragallah, Development of serious games for teaching information security courses, in: IEEE Access, vol. 7, 2019, 169293–169305, https://doi.org/10.1109/ACCESS.2019.2955639.

[14] S. Nicholson, A recipe for meaningful gamification, in: Gamification in Education and Business, Springer, New York, NY, USA, 2015, 1–20, https://doi.org/10.1007/978-3-319-10208-5_1.

[15] S. Papert, Mindstorms: Children, computers, and powerful ideas, Basic Books, New York, 1980.

[16] S. Rajeev, A. Sahu and S. Sharma, Game theme based instructional module to teach binary trees data structure, in: Proc. ISCA 26th Int. Conf. on Software Engineering and Data Engineering, 2017, 13–18, https://sharadonly.github.io/papers/19.pdf.

[17] S. Rajeev and S. Sharma, Educational game-theme based instructional module for teaching introductory programming, IECON 2018 – 44th Annual Conference of the IEEE Industrial Electronics Society, Washington, DC, USA, 2018, 3039–3044, https://doi.org/10.1109/IECON.2018.8592835.

[18] S. Rajeev, S. Sharma and J. Stigall, Game–Theme based instructional module for teaching object-oriented programming, in: Proc. IEEE Int. Conf. on Computational Science and Computational Intelligence (CSCI), 2015, 252–257, https://doi.org/10.1109/CSCI.2015.35.

[19] P. Rule and V. M. John, Your guide to case study research, Van Schaik Publishers, Pretoria, South Africa, 2011.

[20] J. R. Savery, Overview of problem-based learning: Definitions and distinctions, Interdiscip. J. Probl.-based Learn. 1(1) (2006) Art. 3, https://docs.lib.purdue.edu/ijpbl/vol1/iss1/3/.

[21] S. Sharma and E. Ossuetta, Virtual reality instructional modules in education based on gaming metaphor, in: Proc. IS&T Int. Symp. on Electronic Imaging: The Engineering Reality of Virtual Reality, 2017, 11–18, https://doi.org/10.2352/ISSN.2470-1173.2017.3.ERVR-090.

[22] K. Squire, Cultural framing of computer/video games, Game Studies, 2(1) (2002), https://gamestudies.org/0102/squire/, retrieved 2 October 2023.

[23] J. S. Stigall and S. Sharma, Virtual reality instructional modules for introductory programming courses, in: Proc. IEEE Integrated STEM Education Conf. (ISEC), Princeton, NJ, USA, 2017, 33–41, https://doi.org/10.1109/ISECon.2017.7910245.

[24] S. Woods, Loading the dice: The challenge of serious video games, Game Studies, 4(1) (2004), https://www.gamestudies.org/0401/woods/, retrieved 2 October 2023.

Received on March 12, 2023
Accepted on May 7, 2023

Mitko Trpkoski, Andrijana Bocevska,
Igor Nedelkovski and Snezana Savoska

Faculty of Information and Communication Technologies
University "St. Kliment Ohridski"
Boulevard 1st of May bb
Bitola, 7000
REPUBLIC OF NORTH MACEDONIA

E-mails: trpkoski.mitko@uklo.edu.mk
andrijana.bocevska@uklo.edu.mk
igor.nedelkovski@uklo.edu.mk
snezana.savoska@uklo.edu.mk

ГОДИШНИК НА СОФИЙСКИЯ УНИВЕРСИТЕТ „СВ. КЛИМЕНТ ОХРИДСКИ“
ФАКУЛТЕТ ПО МАТЕМАТИКА И ИНФОРМАТИКА
Том 110

ANNUAL OF SOFIA UNIVERSITY “ST. KLIMENT OHRIDSKI”
FACULTY OF MATHEMATICS AND INFORMATICS
Volume 110

NEXT GENERATION SERVICE MODELS OF MOBILE AUTONOMOUS ASSISTANTS

ELITSA VENCHOVA and VASIL GEORGIEV

In this paper, we present an evolution overview of the autonomous streaming-based services for mobile autonomous assistants in the human environment. In Section 1 we present the hypothesis that the technologies are in transition from fixed or portable and wearable devices to those in which the main function is based on independent and possibly self-initiated and remote-controlled movement of the devices. This transition is outlined and categorized in Section 2. Retrospectively, with the progress of the penetration and technologies of the fixed devices, they have evolved from those that function separately and in isolation to those that offer group services in networks of multiple devices known today as Internet of Things (IoT). Another point of our considerations is that consequently and analogously, mobile autonomous devices are also subject of evolution towards service models of multiple cooperating mobile devices. We propose a multidimensional and layered taxonomy for the basic case of isolated mobile autonomous systems for 1d-, 2d-, and 3d-movement models. Our classification covers both the functional and technological aspects of these systems as well as the possibility for grouping services. In Section 3 we consider the parameters of two exemplary platforms as use cases for the two movement modes. Although they represent different principles of movement (i.e. legged and wheeled movement) these examples clearly show a common pattern of service parameters. The movement pattern presented here includes space and time range, speed, range of sensory monitoring, as well as parameters characterizing obstacle surmountability. Finally, in the Conclusion we address the applicability of our parametrization for the purposes of standardization and interoperability the main conditions for the service implementation based on a scalable group of moving devices.

Keywords: Internet of Things (IoT), Autonomous and Intelligent Systems (AIS), mobile autonomous assistants, sensor networks, assistive technologies, wireless integrated network sensors

CCS Concepts:

• Computer systems organization~Embedded and cyber-physical systems~Robotics~Robotic autonomy

DOI: 10.60063/GSU.FMI.110.139-150

1. The transition from fixed to mobile autonomous assistance

The monitoring of various context-specific parameters of the environment using appropriate sensors has passed through two stages and is about to enter a new third stage (or generation). In the first two stages, the devices are static unless their primary function involves passive relocation – typically when embedded in a vehicle or serving a wearer.

First stage is the generation of the Deeply Embedded Systems. These are the smart devices or all specialized devices that perform their limited functions autonomously or semi-autonomously. They do not use network connectivity, or if used, it is to provide a remote user or control interface. Example: an air conditioner that performs its functions according to a daily or weekly program, but with the possibility of distant control e.g. via WiFi protocol.

Second stage – Distributed Embedded Systems widely known as Internet of Things [4]. In it, smart devices, in addition to some local functionality, provide functionality for accumulating and processing sensor data from nodes that are connected to the Internet (or another network that may not be global, but private or limited in some appropriate range).

A major disadvantage of the autonomous assistants of the first and second stages is that the sources of observational sensor data are static or fixed. This has significant negative consequences, the most obvious of which are in two directions:

1. the need for expensive and, in most cases, incompatible and irreplaceable embedding of fixed devices with appropriate saturation in an already built infrastructure;

2. limited functionality – for example, limited possibilities for remote control due to the complex embedding of static systems with actuators for influencing the environment.

By the emerging new generation of autonomous assistants with surface or even airborne and submersible mobility, an important opportunity is created to monitor the environment and infrastructure without having to modify the already built infrastructure. Furthermore, there is no need to saturate the environment with numerous static devices. Instead, the autonomously moving devices can perform their services by moving from point to point and thus covering a wider area. A very obvious example of this transition is the possible replacement of the old-fashioned but widely spread C.C.T.V. (Closed Circuit TV cameras) with a single or several electronic “dogs”. In this use case, the static cameras for an area surveillance are needed to be of enough number to cover the whole area with all the vulnerable spots. Another requirement in C.C.T.V. is that each camera has to be constantly monitored by another camera in the group. Replacement of the static cameras with the moving surveillance assistants may upgrade this functionality by dynamic context-driven repositioning of the registering cameras on and even off the ground.

Depending on the service context the mobile autonomous assistants may have to perform certain humanoid functions by robotic arm gestures in order to control rudimentary existing environment components such as electrical power or control switches or push-button devices. In other words, mobile autonomous assistants are platforms with good potential for adding appropriate actuator components such as a robotic hand for one- and two-finger gestures, which enables the control of various mechanical devices already existing in the environment, such as electrical and electronic switches, buttons, sliders, etc.

Our main objective is to study and analyze the approaches of mobile autonomous monitoring assistants in different application scenarios. The services supported by such groups of multiple moving sensors data monitors can be integrated with traditional and existing fixed monitoring infrastructure.

2. A taxonomy of mobile autonomous assistants

In our taxonomy, the main division of autonomous systems is whether they are fixed including portable or mobile by self-initiative. In mobile autonomous systems, it plays a major factor in the implementation of the assistant to be able to react to surrounding events including potential obstacles but also other context defined events. When the environment is un-urbanised, the moving assistant is expected to be able to navigate through any terrain in its context, while in an urbanized environment, which is more structured, it is expected to be able to handle sidewalks, stairs, doorways, etc. according to whether it is operated in open or closed space. Application in urban environments can be divided into two types – application in public spaces such as airports (indoor), offices, hospitals, museums, schools, etc. or used for private needs at homes. It is obvious that the moving assistants cannot provide logical and mechanical support for any type of terrain and obstacles. E.g. a home assistant can be easily implemented by upgrading a robotic vacuum cleaner if only its operation scenario is for single-floor premises or in buildings with available elevators. Climbing stairs option is more expensive at least in mechanical aspect. Same type of distinction between the types of moving devices can be made for outdoor scenarios.

On the other hand, despite of the vast variety of use cases and surface scenarios we have very limited movement models. Most obvious mechanical models of movement are based on some of the few options. Our taxonomy identifies four categories of several options each listed as follows. A visual representation of our taxonomy is shown in Figure 1.

A. Movement models:

1. "Legs" based walking
2. Wheels based: for plain surface
3. Stair climbing wheels triad: for stairs and stair walks of limited height
4. Track chains (typically of steel or plastic) or track belts: for any unstructured outdoor terrain with parameters reflecting limits of unevenness (fixed obstacles' size and slope angle) and ground material instability (load capacity)

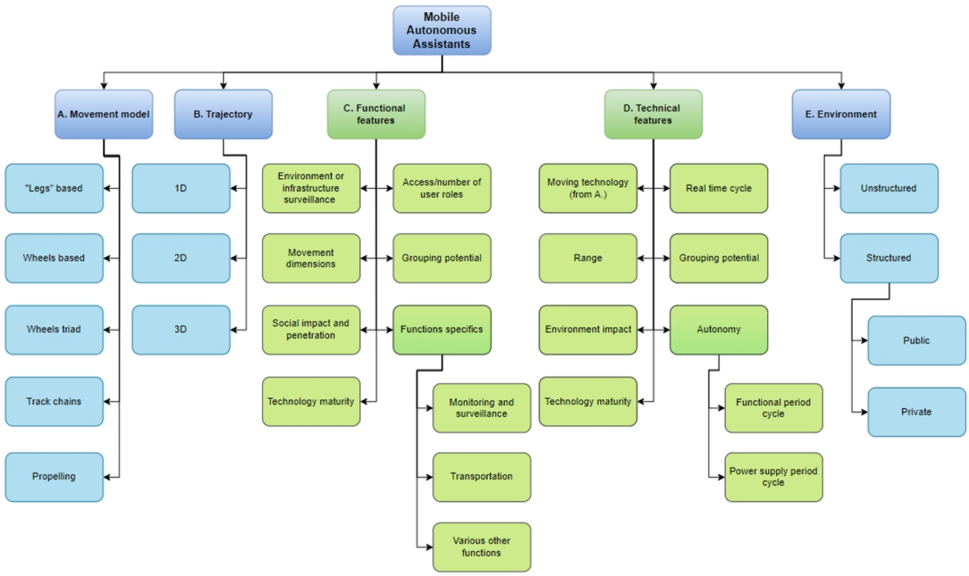

Figure 1. A mobile autonomous assistants

5. Propelling: for any type of 3d movement in fluids (water or air environments) with parameters reflecting limits of the currents and winds

B. Trajectory models:

1. 1d – borders or perimeters
2. 2d – surface
3. 3d – airborne or submerging environmental volumes

C. Functional features:

1. Environment or infrastructure surveillance
2. Access/number of user roles
3. Movement dimensions
4. Functions specifics
   4.1. Monitoring and surveillance
   4.2. Transportation
   4.3. Various other functions (e.g. vac, lawnmower)
5. Social impact and penetration (including useful and harmful)
6. Grouping potential – measured by the scalability (1 for single operating devices)
7. Technology maturity in sense of functionality support, QoS, user experience, user perception and social penetration

D. Technical features:

1. Moving technology – walking, wheeled, propelling, jet, etc. – from A.
2. Autonomy
   2.1. Functional period cycle – time complexity of tasks to perform without user interface
   2.2. Power supply period cycle
3. Range – dimensional range and time to leave (power)
4. Real time cycle – human capacity of perception is 10 Hz. The mobile assistants of course may perform at much higher rates.
5. Environment impact – mechanical impact/digital streaming
6. Grouping options
7. Technology maturity in sense of availability including Price

3. Implementations of mobile autonomous assistants

Here we present just two of the possible use cases of moving autonomous assistants in order to consider their major taxonomy feature A. – the movement model on solid surfaces. By these two cases, we justify the usage of a wheel-based drive (of the cases A1, A2, and A3) to that of the walking "legs"-based robots.

3.1. ANYmal

Robots using legs have an advantage in flexibility and mobility over those using chains and wheels. Despite these advantages, human-like robots are still very far from achieving natural movements. These robots are slow, and require a lot of energy and computing power. Much greater efficiency can be achieved by using multiple legs. The biggest example is Boston Dynamics' Big Dog and Spot. Similar examples in terms of mobility, dynamics and maneuverability can also be seen in the development of the IIT's four-legged hydraulic drive HyQ and the next HyQ2max, MIT's electric cheetah and the ETH's StarlETH serial elastic robot [1].

Developed by Swedish company ANYbotics, ANYmal is a four-legged robot designed for use in an industrial environment to perform inspection and maintenance tasks. It can operate in challenging environments such as stairs, uneven surfaces and confined spaces. The robot is equipped with a set of sensors for navigation and perception [5]. Unlike the examples in the previous paragraph, which were mostly considered in a laboratory environment, ANYmal is one of the first robots used in a real environment [1]. Figure 2 shows the main features of ANYmal.

ANYmal is fully protected against dust and water (IP67), which allows it to operate in humid and dusty environments, saving people from unnecessary exposure to dangerous situations, for example, areas with a high probability of an explosion.

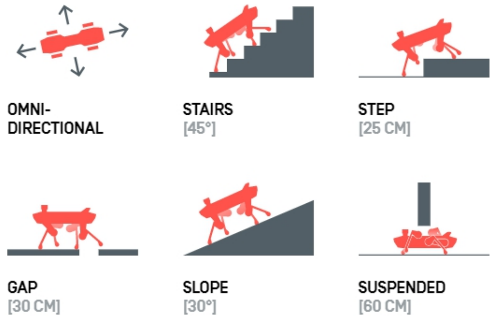

Figure 2. ANYmal characteristics [1]

For this case, ANYbotics also develops Ex-proof ANYmal for potentially explosive environments. Additional protection from water allows it to be easily washed.

It autonomously navigates in a complex environment by having prior information about the environment and finding the most direct way to complete the task. There are built-in depth sensors for precise obstacle avoidance for smoother navigation. Accuracy up to 1cm in both closed and open spaces [1].

Once taught where to go and what to do, he repeats regular inspections on his own. If manual intervention is required, ANYmal allows manual control and viewing from the robot's front and rear cameras. The camera is capable of up to 20× optical zoom to capture clear images and video at long distances in 2K resolution. The user interface allows showing the location of the robot, mission progress and more. At any time, the mission can be interrupted, live camera data can be transmitted, movements can be controlled and additional measurements can be taken for greater precision [1].

After completing a mission, it returns to the charging station. The battery capacity is 90 min. It takes 100 min to quickly charge up to 70%, and 3 hours to fully charge. To extend the range of operation, several charging stations can be installed in different locations. When it runs out of power, it will autonomously go to the nearest docking station to recharge, then the mission will continue [1].

Local data processing (Edge Computing) reduces response time and saves network load. It also eliminates the need for continuous network connectivity. ANYmal has a built-in option to connect to Wi-Fi if available or build his own local network. Through an additional module, it enables 4G/LTE telecommunication. The robot collects and stores the verification data and when the network is available, it will encrypt and feed the information to the ANYmal API and other systems [1]:

- Results can be integrated into Enterprise Asset Management (EAM) or Maintenance Management System (CMMS) management and maintenance systems. Inspection data and analysis reports are transmitted.
- Provide complete reports in PDF or XML format for quick decision making.
- Time and geographic inspection observations.

- Track patterns and historical issues.
- Access to necessary inspection data for further analysis.

The ANYmal data workflow diagram is shown in Figure 3.

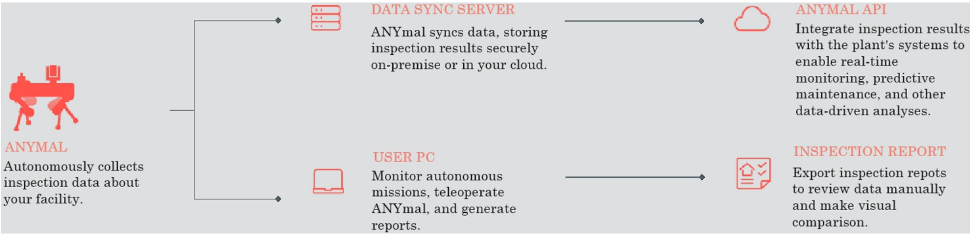

Figure 3. ANYmal workflow [1]

During an inspection, ANYmal provides visual, thermal and acoustic information about the condition of the equipment and the environment. The Pan-tilt module is used to scan the surrounding environment and precisely position the built-in survey sensors to the target point, which helps when it is far away or in a hard-to-reach place. Algorithms based on artificial intelligence analyze the data from the sensors, interpret the values, classify the result and detect anomalies. Unforeseen technical events can lead to dangerous situations – through sensitive sensors it can detect dangerous conditions in the environment and trigger a warning when necessary. Early signs of operational problems are caught by examining the general condition of the equipment. During inspections, ANYmal checks critical points for anomalies and immediately reports serious problems. Industrial environments require continuous monitoring for structural changes. ANYmal scans and documents the data from the 3D environment as a digital twin. All inspection data is linked to accurate temporal and 3D spatial information [1].

Some of Animal's components are: thermal camera provides precise measurements in the range from -20 to $500\,^{\circ}$C without the need for physical contact; LED spotlight aids visual inspections in low- or no-light environments; microphone for recording acoustic measurements in the sonic and ultrasonic frequency range [1].

A wide range of analog instruments and indicators can be digitized. Once trained on a given tool type, the value or state is reliably identified. In case the object is moved to a different angle, the robot can be programmed to take a new position to accurately capture the data [1].

ANYmal allows it to be programmed with additional modules to extend its capabilities through ROS (Robot Operating System) APIs and an open-source ecosystem. A simulation environment that provides a realistic simulation of the physical environment and access to all sensors and APIs facilitates development and testing. For example, a CAD model of the real environment can be loaded and realistic simulations of inspection missions, operating in hard-to-reach places and software integration can be performed before the robot is released into a real environment.

Another option is to walk the robot through the facilities and photograph the inspection points, thereby planning a mission on site. The robot will remember the tasks and independently repeat the inspections [1].

Pharos is a proprietary 3D SLAM software that provides [1]:

- Mapping accuracy of 1–3 cm by fusion of Lidar and depth camera data.
- Full coverage of large facilities with the ability to build maps with up to 4 km range at once.
- Continuous operation without the need for navigation markers, QR codes and GPS connectivity.

Trekker is an artificial intelligence-based deep learning algorithm for navigating complex environments. It allows the robot to access hard-to-reach places without the need to change the environment. Allows ANYmal to climb stairs, scale steps, and crawl over obstacles. Provides performance and reliability. Optimizes the path of routine monitoring tasks [1].

ANYmal constantly checks for people and objects in its surroundings to avoid collisions. It can also move moving obstacles, wait for a path to clear, or find an alternate path to its final destination. There is an option to work together with a fleet of ANYmals to more easily expand coverage and scan frequency [1].

In summary, ANYmalis used for inspection and maintenance in challenging environments. Its main features are quadrupedal design, range of sensors and cameras, and ability to traverse stairs and uneven terrain. As advantages can be marked as high mobility, the ability to operate in challenging environments and a range of sensing capabilities. Some of the ANYmalis limitations are that it has limited load capacity and it is specially designed for inspection and maintenance.

3.2. Jackal

Jackal is a four-wheeled robot developed by Clearpath Robotics for use in research and industrial environments. It can operate both indoors and outdoors. It has a modular design and it is equipped with sensors for precision and control [2].

The Jackal robot is equipped with an integrated onboard computer, GPS, and IMU that work together with ROS to provide a wide range of autonomy options. It can connect via Bluetooth and Wi-Fi and has options for adding various types of sensors with power supply options of 5 V, 12 V, and 24 V. The robot has a sturdy aluminium chassis with a high-torque 4×4 drive that makes it suitable for all-terrain operation. It is weatherproof with an IP62 rating and can operate within a temperature range of $-20\,^{\circ}$C to $45\,^{\circ}$C. Figure 4 shows Jackal human-machine interface [2].

Clearpath Robotics partnered with NVIDIA to develop the Jackal robot, which is built around the Jetson AGX Xavier computing device. The Jetson platform is ideal for robot development with its powerful and compact GPU, making it suitable

Figure 4. Jackal HMI [2]

for VSLAM, 3D imaging, and machine learning applications. The robot is configured for basic autonomous operations in both indoor and outdoor environments, from GPS waypoint navigation with laser scanning for collision avoidance to indoor mapping and route planning. It comes with a pre-installed Linux and ROS system [2].

The visual survey setup of the Jackal robot includes two front-facing FLIR Blackfly cameras for stereo vision and is fully compatible with ROS, RViz, and Gazebo. It also has a pre-installed CUDA toolkit, making it suitable for machine vision, human-machine interaction, and Visual SLAM applications [2]. Gazebo simulation packages enable testing of the robot's capabilities in a virtual environment, enabling early detection of problems and gaps. Gazebo provided an almost identical configuration to the actual Jackal, and various packages, services, nodes, and themes were available in the simulation [5].

The navigation package of the Jackal robot combines a SMART-7 RTK GPS with a wide-range station, achieving an accuracy of less than 2 cm and remote communication up to 1 km. This package is used for outdoor navigation, multi-robot systems, and remote environmental sensing [2].

Many additional extensions such as lidars, cameras, GPS systems and many more can be added. For example [2]:

- SICK TIM551 LiDAR
- 3DM-GX3-25 IMU
- TIM551 laser range finder
- SwiftNav Duro GPS
- IP65 FLIR Camera Enclosure
- SMART-6 RTK
- HDL-32e 3D laser scanner

In summary, Clearpath Jackal is used for research and industrial applications. Its main features are a four-wheeled design, the ability to operate in outdoor and indoor environments, a modular design, and a range of sensors and cameras. As advantages can be marked as high mobility, the ability to operate in challenging environments, a range of sensing capabilities and modular design. Some of Jackal his limitations are that it has limited load capacity and it is specially designed for research and industrial tasks.

3.3. Comparison

Table 1 compares some of the technical characteristics of ANYmal and its components with those of Jackal. Comparison is divided into different categories – ANYmal and Jackal, Payload, Perception Sensors, Communication, Environment, Battery and Battery Charger.

Table 1. Comparison of technical characteristics [1,2]

	ANYmal	Jackal
Dimensions L × W × H	930 × 530 mm (default walking) × 890 mm (default walking) / 470 mm (lying on the ground)	508 × 430 × 250 mm
Weight	50 kg / 55.7 kg with Inspection Payload	17 kg
Speed	1.3 m/s maximum, rough or slippery terrain may reduce the walking speed, 0.75 m/s recommended for safe and efficient operation	2.0 m/s
Number of limbs/ wheels	4	4 × 190 mm diameter
Degrees of freedom	12	
Payload		
Weight	5.7 kg	all terrain: 10 kg maximum: 20 kg
Perception Sensors		
LIDAR	16 channels, 300 000 points/s, full sweep at 10 Hz 0.4–100 m range, 3 cm accuracy (typical) 360 × 15.0 to −15.0° FOV (Horizontal × Vertical) 905 nm, Class 1 Eye-safe per IEC 60825-1:2007 & 2014	SICK TIM551 LiDAR
Depth camera	0.3–3 m range, 87.3 × 58.1 × 95.3° depth FOV (Horizontal / Vertical / Diagonal), Class 1 Laser Product under the EN/IEC 60825-1, Edition 3 [2014]	IP65 FLIR Camera Enclosure
Tele-operation cameras	1440 × 1080 px 110 × 76.5 × 117.7° FOV (Horizontal / Vertical / Diagonal)	
Internal Sensing		Battery Status, Wheel Odometry, Motor Currents, Onboard IMU, Onboard GPS
Communication		
Communication	Wi-Fi: Built-in module 2.4/5 GHz, 802.11 ac wave2 Access point or client mode 4G LTE: Add-on module, LTE Cat.12	Ethernet, USB 3.0, RS232, IEEE 1394 avail.
Environment		
Temperature	Specified: 0–40 °C Typical: −10–50 °C	−19 to 45 °C
Water & dust ingress protection	Fully protected against water and dust (IP67) and able to operate in humid and dusty conditions.	Designed for IP62

Battery		
Battery & capacity	Swappable Li-ion battery, UN 38.3 certified 932.4 Wh	Lithium, 270 Wh
Running time & range	90–120 minutes on a full charge 4 km range on a full charge, up to 2 km for a typical inspection mission depending on payload weight and number of inspection points.	2 hrs maximum 8 hrs typical
Recharge time	3 h for a full charge, 100 min for a 70% quick charge	4 hours
Weight	5.5 kg	
Battery charger		
Power supply	110–240 V/50–60 Hz	110–220 VAC

As can be seen from the overview above and Table 1, one of the main differences is that the battery capacity of the first use case (ANYmal) is about 3.5 times that of the second use case (Jackal) and yet the time the ANYmal's operation time is 90–120 minutes, while the Jackal's is 8 hours in typical use. There is also a difference in speed between the two examples – Jackal's speed is 2.0 m/s, while ANYmal's is 1.2 m/s, but the recommended 0.75 m/s for safe and efficient operation is most likely influenced by a more complex motion model of ANYmal and its 12 degrees of freedom. Another factor in the result above is the big difference in weight, where the ANYmal weighs 50 kg without Inspection Payload, and the Jackal weighs 17 kg. The time to fully charge the first use case and the second is approximately the same – 3 hours and 4 hours, respectively.

Due to their roughly the same application areas (indoor and outdoor inspection), the two examples have fairly similar location and environmental analysis components – GPS, remote controller (via Wi-Fi or Bluetooth), depth sensors and cameras. Both use cases allow to be programmed with additional modules to extend their capabilities through ROS APIs and an open-source ecosystem and both have simulation environments for easier and better development.

4. Conclusion

This paper presents a taxonomy of the autonomous streaming-based services based on self-initiative moving devices. We are focusing on the transition from fixed or portable and wearable devices to those with independent and self-initiated movement. The taxonomy provides a multidimensional and layered classification (movement, environment and functionality) for isolated mobile autonomous systems with 1d-, 2d-, and 3d-movement models, covering both functional and technological aspects. The paper also considers two exemplary ground moving platforms as use cases for the movement models based on walking and wheels – ANYmal and Jackal. Our concept which is reflected in the taxonomy is that the usage of a wheel-based drive (of the cases A1, A2, and A3) has very important advantages to that of the walking "legs"-based robots. Overall, this paper contributes to the understanding of the evolution of autonomous streaming-based services and provides a framework for further research and development. In a more general point of view currently the artificial moving assistants mimic in their evolution that of the living creatures but only with some important differences. Wheel drive is one of the most important among them. Of course, in the end, this reflects the difference between a proteins-based civilization and the artificial one based on silicon, metals and mathematics.

Acknowledgements

This paper is prepared with the support of MIRACle: Mechatronics, Innovation, Robotics, Automation, Clean technologies – Establishment and development of the Center for Competence in Mechatronics and Clean Technologies – Laboratory Intelligent Urban Environment, funded by the Operational Program Science and Education for Smart Growth 2014–2020, Project BG 05M2OP001-1.002-0011.

References

[1] ANYbotics, Generation D ANYmal Technical Specifications, https://www.anybotics.com/wp-content/uploads/2022/09/anymal_specifications_sheet.pdf. Accessed 12 April 2023.

[2] Clearpath Robotics Inc., Jackal UGV Tutorial, Jackal Tutorials, Clearpath Robotics Inc., 2015. [Online] https://docs.clearpathrobotics.com/docs/robots/outdoor_robots/jackal/user_manual_jackal/. Accessed 12 April 2023.

[3] M. Hutter et al., ANYmal – toward Legged Robots for Harsh Environments, Advanced Robotics 31(17) (2017) 918–31, https://doi.org/https://doi.org/10.1080/01691864.2017.1378591.

[4] International Telecommunication Union (ITU), Internet of Things Global Standards Initiative, https://www.itu.int/en/ITU-T/gsi/iot/Pages/default.aspx. Retrieved 28 July 2022.

[5] A. Ollman, Implementation of Pose Estimation Algorithms on the Clearpath Jackal UGV, 2019.

Received on April 14, 2023
Accepted on May 7, 2023

Elitsa Venchova and Vasil Georgiev

Faculty of Mathematics and Informatics
Sofia University "St. Kliment Ohridski"
5 James Bourchier Blvd.
1164 Sofia
BULGARIA

E-mails: evenchova@fmi.uni-sofia.bg
v.georgiev@fmi.uni-sofia.bg

Submission of manuscripts. The *Annual* is published once a year. A `pdf` file of the manuscript should be sent to the e-mail address of the Editorial Board `annuaire@fmi.uni-sofia.bg`. Once received by the editors, the manuscript will be subjected to rapid, but thorough review process. If accepted, it is immediately scheduled for the nearest forthcoming issue. No page charge is made.
The submission of a paper implies that it has not been published, or is not under consideration for publication elsewhere. In case it is accepted, it implies as well that the author(s) transfers the copyright to the Faculty of Mathematics and Informatics at the Sofia University "St. Kliment Ohridski", including the right to adapt the article for use in conjunction with computer systems and programs and also reproduction or publication in machine-readable form and incorporation in retrieval systems. A printable pdf file of the published paper will be available on the journal website.
Instructions for authors. Preferences will be given to papers, not longer than 25 pages, written in English and typeset by means of LaTeX or MS Word. MS Word files will be converted to LaTeX. Templates are available on the journal website.
Manuscripts should be submitted for editorial consideration in `pdf` format by e-mail to `annuaire@fmi.uni-sofia.bg`. Upon acceptance of the paper, the authors will be asked to send the text of the papers in `tex` or `docx` format and the appropriate graphic files of the figures (if any).
The first page of manuscripts must contain a title, name(s) of the author(s), a short abstract, a list of keywords and appropriate codes according to Mathematics Subject Classification 2020 (MSC2020), see `https://msc2020.org/` or 2012 ACM Computing Classification System (CCS), see `https://dl.acm.org/ccs`. The affiliation(s), including the electronic address, should be given at the end of the manuscript.
Figures have to be inserted in the text near their first reference. The graphic files in any format of the following `jpg`, `png`, `pdf`, `eps` should be submitted together with the final version of the manuscript after acceptance. All illustrations are to be in gray scale or black-and-white, else appropriate conversion will be done.
Tables should be inserted in the text as close to the point of reference as possible. Some space should be left above and below the table.
Footnotes must be numbered consecutively and should be brief.
References must be cited in the text in square brackets, like [3], or [5, 7], or [11, p. 123], or [16, Ch. 2.12]. They have to be numbered alphabetically by the family name of the first author. Examples (please note order, style and punctuation of the bibliographical description of each entity):
For books: R. R. Phelps, Convex functions, monotone operators and differentiability, Second Edition, Lecture Notes in Mathematics, 1364 Springer-Verlag, Berlin, 1993.
For journal articles: J. M. Borwein and D. Preiss, A smooth variational principle with applications to subdifferentiability and to differentiability of convex functions, Trans. Amer. Math. Soc. 303(2) (1987) 517–527.
For articles in edited volumes or proceedings: R. Deville and N. Ghoussoub, Perturbed minimization principles and applications, in: Handbook of the geometry of Banach spaces, ed. by W. B. Johnson and J. Lindenstrauss, vol. 1, North-Holland, Amsterdam, 2001, 393–435.