NAM CAN THO UNIVERSITY

Số 25
2023
ISSN 2588 - 1272

Tạp chí
KHOA HỌC & KINH TẾ PHÁT TRIỂN

INFORMATION TECHNOLOGY

JOURNAL OF SCIENCE AND DEVELOPMENT ECONOMICS
NAM CAN THO UNIVERSITY

TABLE OF CONTENTS

Pages

Journal of Science and Development Economics
Nam Can Tho University

Website: jsde.nctu.edu.vn

Needs and solutions for digital transformation in the tourism sector in Can Tho City

Dao Ngoc Canh[1*], Nguyen Van Linh[2], Tran Thanh Nam[2], Nguyen Du Ha Long[1], Doan Hoa Minh[2], Tran Van Thien[2]

[1]*Faculty of Tourism and Hospitality Management, Nam Can Tho University*

[2]*Faculty of Information Technology, Nam Can Tho University*

[*]*Corresponding author: Dao Ngoc Canh (email: dncanh@ctu.edu.vn)*

Received: November 10, 2023
Revised: November 30, 2023
Accepted: December 5, 2023

Keywords: *Can Tho City, digital transformation, smart tourism, virtual reality*

Từ khóa: *chuyển đổi kỹ thuật số, du lịch thông minh, thành phố Cần Thơ, thực tế ảo*

ABSTRACT

In today's era, the fourth industrial revolution, coupled with the rapid advancement of information and communication technology, has led to a shift towards a digital platform in all aspects of human activities. As a result, digital transformation has become an inevitable and crucial trend for organizations, individuals, and businesses to survive and thrive. This is especially true for the tourism industry, where the need for quick and effective digital transformation to meet customer demands and enhance international cooperation has become urgent. This article discusses the necessity of digital transformation in the tourism industry in the context of the fourth industrial revolution and proposes specific solutions to promote digital transformation in Can Tho City's tourism sector.

TÓM TẮT

Trong thời đại ngày nay, cuộc cách mạng công nghiệp 4.0 cùng với sự phát triển vượt bậc của công nghệ thông tin và truyền thông làm cho mọi mặt hoạt động của con người đều chuyển sang nền tảng số. Vì vậy, việc chuyển đổi số trở thành xu hướng tất yếu, vô cùng quan trọng và cần thiết đối với tất cả các tổ chức, cá nhân và doanh nghiệp để tồn tại và phát triển. Đặc biệt, đối với ngành Du lịch, vấn đề chuyển đổi số nhanh chóng và hiệu quả để đáp ứng nhu cầu khách hàng và tăng cường hợp tác quốc tế đang trở nên cấp thiết. Bài viết này đề cập đến nhu cầu chuyển

đổi số Du lịch trong bối cảnh cuộc cách mạng công nghiệp 4.0 và đề xuất một số giải pháp cụ thể góp phần thúc đẩy công tác chuyển đổi số của ngành Du lịch tại thành phố Cần Thơ.

1. INTRODUCTION

The strong development of information and communication technology requires all activities to move to digital platforms. Therefore, digital transformation has become an inevitable trend, extremely important and necessary for businesses and individuals to survive and develop. “Digital transformation is about rethinking how organizations bring together people, data and processes to create new value” (Microsoft, 2017) [1]. To put it simply, digital transformation means moving from a face-to-face working environment to an online (digital environment) with more data and connections. For example, thanks to the support of digital technology, technology taxi companies such as Grab, Uber... or online businesses are operating in new business methods. They may not own any cars or houses but can rent cars and houses all over the world. For the tourism industry, digital transformation is the process of applying digital technology to business activities and tourist experiences, completely changing the tourism model from traditional to modern. From there, "traditional" activities such as storing documents (personnel list, tour list, customer list etc.) or marketing activities (distributing leaflets, posting advertisements, etc.) will be done using digital technology, thereby making business administration more effective, economical and faster.

Digital transformation also enhances tourist experience activities along the digital value chain. Visitors will experience convenient and impressive resort and entertainment services. In addition, digital transformation also changes marketing methods, creating online interaction channels between tourists and businesses. The development of the Internet and electronic devices such as mobile phones and computers with network connections, especially wireless networks (wifi), makes it easy for people to find diverse and attractive sources of information, and create travel inspiration through websites or social networks. The smart tourism model also provides suggestions on tourist destinations for guests to choose from and stimulates travel inspiration, while also supporting guests in making a quick and effective trip. However, to meet the needs of digital transformation in the tourism industry, there needs to be coordination from many parties, from researchers, managers, tour guides, marketers, tour operators to tourists and service providers of accommodation, food, transportation and so on. This article analyzes the overall need for digital transformation in tourism in the context of the 4.0 industrial revolution and proposes a number of solutions to contribute promoting digital transformation of the tourism industry in Can Tho City.

2. METHODS

In this study, the researchers mainly used secondary data collection and analysis methods. Data were collected from research projects, statistics and summary reports of Can Tho City. These data were analyzed and synthesized to serve the research topic. At the same time, the research team also conducted consultation with experts including scientists, tourism industry

managers and tourism businesses in Can Tho City.

3. RESULTS

3.1 Some issues about digital transformation in tourism

3.1.1 Digital transformation in tourism

Previously, traditional tourism business mainly focused on the intrinsic value of the business, with businesses and tourists transacting directly with each other. The advent of the Internet has led to the stage of e-commerce. During this period, the enterprise's internal business activities have been supported by information technology and have developed further, enhancing the connection between businesses and tourists through online transactions: business and marketing activities via the Internet. The current digital transformation period is a new step of development compared to the previous period, notably the smart tourism model with the ability to connect and use services with "smart" elements to offer. Assessments for businesses about potential tourists, predicting tourists' needs and preferences. On the other hand, tourists can also enhance their travel experience with smart services such as suggestion systems, real-time systems, online connections, etc.

Optimizing user experience and business operations by allowing customers to seamlessly access diverse data to understand products and services will help businesses meet customer requirements. in the most convenient and time-saving way. On the basis of digital transformation, modern tourism business models and online travel agency channels with the appearance of a series of "big players" such as Booking, Agoda, Expedia, Traveloka, Klook, ect., shows that the tourism market is extremely competitive and new changes and trends that help streamline the apparatus will play a decisive role in the survival of tourism businesses in the future.

3.1.2 Benefits of digital transformation in tourism

The most noticeable benefits of digital transformation for businesses are cutting operating costs, reaching more customers for a longer period of time, and leaders being able to make timely and accurate decisions. more accurate thanks to a timely and transparent reporting system, optimizing employee productivity, etc. These things help increase operational efficiency and enhance the competitiveness of organizations and businesses. Smart tourism helps interaction and close connection between management agencies, businesses and tourists, to improve the quality of tourist service, and at the same time help businesses operate more effectively. Smart destinations and tourists are core issues in the smart tourism concept. Thanks to a large amount of data on information about tourist destinations and tourism businesses, tourists can easily choose the travel method that suits them and experience interesting forms of tourism. improve product quality of the tourism industry. In addition, smart tourism based on new technology platforms will increase the attractiveness of tourism products, improve efficiency in promoting and marketing tourism products, and contribute to changing tourist behavior. These are important benefits that smart travel brings.

Modern travelers always tend to seek convenience to save time. Therefore, an interactive experience and convenient booking of travel services anytime, anywhere is their desire. They can sit at home but can book

services in advance at any tourist destination they want to visit around the world, from hotel reservations, tour bookings, airline ticket purchases, electronic payments and so on. Thus, tourists just need to "pack their suitcases and go", without even needing to bring cash. That is also a way to change the travel mindset of tourists, contributing to improving the tourist experience. Imagine today if traditional hotels did not have advertising, did not appear on social networks, did not apply new business models, or hotels did not accept electronic payments, would anyone? know or come to stay? At famous tourist destinations where tourists have to stand in line for hours to buy tickets, without an Internet ticket booking system or online ticket sales, how many people are waiting? In fact, there are large tourism businesses in the world that could not adapt to the digital age and had to close, including Thomas Cook Group (considered the "father" of the travel industry) in 2019, after more than 178 years of existence. Meanwhile, many online travel agents (Online Travel Agents - OTA) have grasped the trend and have quickly dominated the market providing travel services through online channels, all transactions from booking services to paying, everything is online. With this lesson, if tourism businesses do not want to be turned away by tourists and go bankrupt, they must carry out digital transformation quickly and effectively.

3.2 New technology platforms serve tourism digital transformation

3.2.1 Big Data

Big data (Jones, 2019) [2] is a term used to refer to very large data sets that cannot be processed or analyzed using traditional methods. Big data has 3 basic characteristics: Volume - Volume of data, Velocity - Speed, data generation speed and data processing speed. Variety – The variety of data from structured, semi-structured and unstructured. Big data in the tourism industry is data about tourist information, tourist needs, habits and behaviors of tourists. This data can be collected from the business itself, originating internally. or outside, from large data warehouses in the world such as Data Warehouse, Data Lake or tourism businesses can provide free services for tourists to use in exchange for information and data from tourists.

3.2.2 Machine Learning

Machine learning (El Naqa & Murphy, 2015) [3] is a branch of artificial intelligence (AI), a field of research that allows computers to improve themselves based on training data sets. Training Data, based on what has been learned, the computer can analyze itself to make predictions or decisions without needing to be specifically programmed. Machine learning will use algorithms to analyze large data in the tourism industry into specific data that is useful for each business, thereby identifying a large amount of information about tourists, travel experiences, reviews, traveler preferences, favorite destinations. The combination of machine learning and big data will benefit tourism businesses in particular and the tourism industry in general by being able to accurately predict future tourist needs and optimize pricing strategies, more precisely targeted marketing and improved visitor experiences.

3.2.3 Internet of Things (IoT)

The Internet of Things (Rose et al., 2015) [4] has the potential to make a huge impact on businesses, automating processes without the need for any human-computer or human-to-

human interaction. People rely on the ability to provide modern connectivity and communication between devices, between systems and services via the Internet. IoT is expected to thrive in the Tourism industry. IoT technology such as tablets in smart hotel rooms will enable personalization to customer needs, such as turning lights on and off, adjusting room temperature, controlling TVs, elevators and air conditioning, schedule an alarm by call. This will bring convenience to travelers similar to their home, making them want to return to the hotel again. Or after a traveler successfully books and pays electronically, the hotel can automatically send an electronic key card to the traveler's smartphone, allowing them to check in without anyone's assistance. Smart locks with NFC readers will ensure security by allowing guests to personally restrict access to amenities as required.

3.2.4 Cloud Computing

Cloud computing technology or virtual server computing is a service model that allows users to easily, anytime, anywhere and on-demand access to shared computing resources (networks, servers, storage) and services) through a network connection. Currently, cloud computing is the application platform of large corporations such as Google, Microsoft and so on. Solutions from cloud technology will help tourism businesses technologize and handle tasks faster with a more professional and coherent process. Communication and collaboration of a tourism business with many branches and offices in many geographical regions is no longer a barrier when using cloud technology. Cloud technology also has the ability to allow thousands of people to collaborate, share data, access information, and make voice or video calls at the same time. With this feature, tourism businesses not only save on training and human resource management costs but also enhance connection, promotion, and information sharing with a series of customers at the same time. Cloud technology also gives businesses the ability to reach tourists, shorten search time and connect with suitable visitors. This will help businesses coordinate, schedule and advise according to the best interests of the business and visitors.

3.2.5 Virtual Reality

Virtual Tour or Interactive Tour has been around since 1994 and has become more popular among tourists in many countries around the world. Virtual tours or interactive tours aim to simulate tourist destinations through images, videos, sound effects, music or reports, introductions, texts, etc. Factors that make virtual tours attractive for tourists, modern technologies such as 360 photos, 360 videos, Panorama photos, Flycams and so on help tourists better understand the place they are about to visit and stimulate inspiration for their travel.

3.3 Current status and solutions for digital transformation of tourism in Can Tho City

3.3.1 Current status of tourism digital transformation in Can Tho city

Can Tho is a centrally- run city, the center of the Mekong Delta region, a focal point connecting the southwestern provinces with the whole country and the world. Can Tho has potential for development in agriculture, industry, trade and services. In particular, Can Tho tourism industry is being invested in and developed, acting as a driving force center to promote the development of the Mekong Delta tourism industry. The Government's decision

approving the Master Plan for tourism development in the Mekong Delta to 2020, with a vision to 2030, has set the goal: "Striving to make Can Tho City become a tourism center and coordinate visitors for the entire the Mekong Delta is one of the tourism development centers of the country." Resolution 10-NQ/TU dated December 29, 2021 of the Can Tho City Party Committee also emphasized: "People must exploit and maximize the city's potential, advantages, and tourism resources, especially the role Regional center, gateway position of the lower Mekong River region associated with promoting the strengths of roads, waterways and airways."

Can Tho has rich tourism resources with a gentle climate, interlaced rivers and canals associated with the typical cultural features of the river region, lush and rich fruit gardens, along with many festivals, folk craft villages, scenic spots, historical and cultural relics and so on. In addition, Can Tho's friendly and hospitable people have created favorable conditions to become an attractive tourist destination and attracting more and more domestic and international tourists to this land rich in tourism development potential.

With its potentials and advantages, tourism activities in Can Tho City have had many changes and achieved positive results. In 2022, the total number of Can Tho tourist arrivals will reach 5,134,605, an increase of 142% over the same period. Accommodation businesses served 2,508,305 visitors, an increase of 179% over the same period. Total tourism revenue is estimated at 4,117 billion VND, up 199% over the same period. In 2023, the tourism industry of Can Tho City sets a goal of welcoming 5,200,000 visitors, with total tourism revenue reaching 4,580 billion VND (Department of Culture, Sports and Tourism of Can Tho City, 2023) [5].

In the general trend of digital transformation worldwide, digital transformation in the tourism industry of Can Tho City has been focused on implementation and achieved positive results, especially in information activities and promoting tourism. In particular, the Can Tho City Tourism Development Center, with its tasks and functions, has become a leading unit implementing digital transformation in tourism. Currently, the Center is managing two electronic information portals that regularly post and update news and articles related to events, tourism information, destinations, and tours in the city. They are tourism information portal: https://tourismcantho.vn with integrated service of Audio Guide technology (automatic voice-over) to help visitors grasp information without reading text and smart tourism information portal with https://canthotourism.vn with a service integrating 3D map technology, through which visitors can view 360-degree images of some tourist destinations. Notably, the smart travel application: "Can Tho Tourism" is run on both Android and IOS platforms to help tourists update and look up travel information in the fastest and most accurate way. In addition, a social network ecosystem including: Facebook, Zalo, Tiktok, Youtube, Instagram "Can Tho Tourism - Can Tho Tourism" is also being built to widely promote the image of people and culture, destinations, and cuisine of the city to domestic and international tourists.

Since 2019, Can Tho City's tourism industry has signed a tourism development cooperation program with Ho Chi Minh City

and 13 provinces and cities in the Mekong Delta, in which digital transformation is one of the key contents. to create motivation to promote tourism development. On July 11, 2023, in Can Tho City, the Digital Transformation Festival in Tourism took place with the theme "Digital Transformation - Driving force for sustainable development". The event once again identified digital transformation as a key factor in creating a breakthrough for tourism links between Ho Chi Minh City and 13 provinces and cities in the Mekong Delta. At the conference, Mr. Nguyen Thac Hien, Vice Chairman of Can Tho City People's Committee, said that for the tourism industry, Can Tho has deployed the Smart Tourism Information Portal and Can Tho City Smart Tourism Application on mobile devices gradually contributing to promoting the effectiveness of its role in supporting the management of tourism activities, connecting people, tourists and businesses. These applications have attracted more than 8 million visitors and interactions, with an average daily visit of more than 7,000 (Ai Lam, 2023) [6].

3.3.2 Some proposed solutions for digital transformation of tourism in Can Tho City

Electronic payment using QR Code:

Electronic payment is one of the features that brings users many conveniences and a convenient and safe experience. Building digital transformation solutions can start from the simplest things like electronic payments. Electronic payment helps tourists use travel services conveniently, flexibly, safely and securely, helping to save time and limit financial risks, contributing to enhancing professionalize tourism in Can Tho City.

Smart travel mobile application:

Smart tourism mobile application uses 4.0 technologies such as Big Data, Machine Learning to provide information about tourist destinations in Can Tho such as scenic spots, restaurants, hotels, and recommendations for visitors to dining and entertainment locations, recommend suitable means of transportation, shop for online specialties with door-to-door delivery, and integrate electronic payment features such as hotel reservations and sightseeing ticket purchase online, automatic check-in, check-out, etc. In addition, through the application, visitors can report tourism businesses, hotels, locations and so on to receive support and guidance on resolving and protecting their legitimate rights, contributing to ensuring a healthy and increasingly civilized tourism business environment.

Figure 1. Smart tourism mobile application in Can Tho City

Building a Semantic tourism website in Can Tho city:

Semantic Web is a new generation of Web, also known as Semantic Web or Web 3.0, helping users find information smarter, faster and more accurately than traditional search engines. Applications related to the Semantic Web allow computers to understand information on the web, support smarter searches, support information extraction, data integration and automate some tasks for people. With the Semantic Web tourism application in Can Tho, visitors can search for tourist destinations according to their needs in the most accurate and effective way, suggesting suitable specific destinations and detailed information about their needs. Destination location such as suggested means of transportation to get there, distance and time to get there, opening times, types of products or business specialties all are searched in detail.

Figure 2. Semantic Can Tho city tourism website

Building tourism business management software for Can Tho City:

For tourism service businesses, tourism business management software supports updating tourism news, serving business operations and management. At the same time, it is software that serves electronic payment and electronic ticket checking quickly and conveniently between travel service businesses and customers - those who use digital platforms.

4. DISCUSSION

4.1 Difficulties of digital transformation in tourism

- *Difficulties in lack of resources:* For businesses conducting digital transformation, they will have to apply new technologies, leading to large initial investment costs, in addition to maintenance costs. Along with the situation where businesses use too much software with separate features, data is not synchronized, costs increase, each unit uses its own software, causing internal communication to become limited.

- *Difficulties in finance and management:* Businesses or localities need a certain investment, both to regularly update the digital transformation system and to train human resources to use the software, digital transformation equipment in tourism. Digital transformation in tourism is still not synchronized between localities. Areas with good conditions for digital transformation in

tourism are mostly in large provinces and cities. Requires more synchronous investment attention from functional sectors. Digitalization activities in the tourism industry are still sporadic, fragmented, and have not been successfully connected and built on a database. This will cause a lack of data (including reports and information analysis), so the process of management, control, reporting and data statistics in the industry faces many difficulties.

- *Difficulties in habits due to business practices:* The majority of Can Tho people in the western region of the river are familiar with the typical traditional tourism model such as direct cash transactions for products and specialties; therefore, wanting to change people's habits is not easy.

4.2 The issue of training for tourism digital transformation

When new technology is born, it will impact human resources, leading to the risk of losing jobs or some training fields will no longer be necessary (for example, if the smart hotel model develops strongly, it will replace for hotel receptionists, travel consulting hotline staff, etc.). Therefore, digital transformation of the tourism industry is also facing difficulties in human resources. To successfully transform digitally, it is necessary to have good human resources and a team of competent experts in both Information Technology and Tourism. However, Vietnam's current training system still has many shortcomings in keeping up with the trend of nurturing talent and developing digital tourism human resources. It will take a lot of time to train a person who is both good at the Tourism industry and has good knowledge of Information Technology. On that basis, a solution proposed by the group is to build a project to open the "e-Tourism" industry with a combination of teaching from lecturers from the Faculty of Tourism and the Faculty of Information Technology to equip students with both knowledge of the Tourism industry and good knowledge of Information Technology after their graduation.

5. CONCLUSION

Digital transformation in general, and tourism digital transformation in particular, have become an inevitable development trend, determining the existence and development of tourism businesses in all countries and localities. In Vietnam, applying information technology to digital transformation is a major policy of the Vietnamese Party and State expressed in many resolutions, policies, programs, plans and so on. In particular, to deploy digital transformation of Tourism industry, the Prime Minister issued Decision No. 1671/QD-TTg dated November 30, 2018 approving the "Master plan for applying information technology in the field of tourism in the period 2018-2020, defining towards 2025" which emphasizes the priority of developing digital tourism and smart tourism. Can Tho City, as a tourism center in the Mekong Delta region, the need for digital transformation in tourism has both intrinsic significance and contributes to promoting the digital transformation process in the Mekong Delta region. The article has outlined a number of solutions as a foundation for digital transformation in tourism in Can Tho City, helping individuals and tourism businesses in Can Tho city to easily access and implement. Digital transformation is faster.

REFERENCES

[1] Microsoft. (2017). *A Strategic Approach to Digital Transformation in Manufacturing Industries.*

[2] Jones, M. (2019). What we talk about when we talk about (big) data. *The Journal of Strategic Information Systems, 28*(1), 3–16.

[3] El Naqa, I., & Murphy, M. J. (2015). *What is machine learning?* Springer.

[4] Rose, K., Eldridge, S., & Chapin, L. (2015). The internet of things: An overview. *The Internet Society (ISOC), 80*, 1–50.

[5] Department of Culture, Sports and Tourism of Can Tho City. (2023). *Report summarizing tourism activities,* 2022.

[6] Ai Lam (2023). *Chuyển đổi số tạo đột phá liên kết du lịch*. https://baocantho.com.vn/chuyen-doi-so-tao-dot-pha-lien-ket-du-lich--a161885.html.

Journal of Science and Development Economics
Nam Can Tho University

Website: jsde.nctu.edu.vn

Application of artificial intelligence in smoke detection with streaming updated data

Tran Ho Xuan Mai[1], Ngo Ho Anh Khoi[1], Trinh The Luc[2], Tran Huynh Khang[1]
[1]*Faculty of Information Technology, Nam Can Tho University*
[2]*Hanoi University of Mining and Geology*

Received: November 10, 2023
Revised: November 30, 2023
Accepted: December 5, 2023

Keywords: *artificial intelligence, decision tree, smoke detection*

Từ khóa: *cây quyết định, phát hiện khói, trí tuệ nhân tạo*

ABSTRACT

At present, artificial intelligence (AI) is one of the most rapidly developing fields in science and technology. In the modern context, AI technologies have become a highly researched area globally, leading to breakthrough technologies that enhance efficiency and effectiveness across various sectors, including environmental security. In today's landscape, where fire and explosion incidents can occur anywhere, there is growing concern about environmental pollution caused by hazardous substances released from fires. These subtainces can come into contact with soil, water, and air, posing a threat to the environment. This technology has the potential to greatly impact human an animal health, as well as the environtment, by reducing the occurrence of diseases and adverse effects caused by fires. To investigate this topic, the "Smoke Detection Dataset" provided by Deep Contractor was selected as the primary dataset. The Decision Tree Classifier algorithm, which has shown significant advancement in various fields of artificial intelligence, was chosen to build classification models for this study. Its proven effectiveness and versatility make it suitable foundation for the system being developed.

TÓM TẮT

Hiện nay, trí tuệ nhân tạo (AI) là một trong những lĩnh vực khoa học công nghệ phát triển nhanh nhất. Trong bối cảnh hiện đại, công nghệ AI đã trở thành một lĩnh vực được nghiên cứu kỹ lưỡng trên toàn cầu, dẫn đến những công nghệ mang tính đột phá giúp nâng cao hiệu suất và hiệu suất trên nhiều lĩnh vực, bao gồm

cả an ninh môi trường. Trong bối cảnh ngày nay, nơi sự cố cháy nổ có thể xảy ra ở bất cứ đâu, mối lo ngại ngày càng tăng về ô nhiễm môi trường do các chất độc hại thoát ra từ hỏa hoạn. Những chất phụ này có thể tiếp xúc với đất, nước và không khí, gây ra mối đe dọa cho môi trường. Công nghệ này có khả năng tác động lớn đến sức khỏe con người và động vật cũng như môi trường bằng cách giảm sự xuất hiện của bệnh tật và tác động bất lợi do hỏa hoạn gây ra. Để điều tra chủ đề này, "Bộ dữ liệu phát hiện khói" do Deep Contractor cung cấp đã được chọn làm bộ dữ liệu chính. Thuật toán Phân loại cây quyết định, đã cho thấy sự tiến bộ đáng kể trong các lĩnh vực trí tuệ nhân tạo khác nhau, đã được chọn để xây dựng các mô hình phân loại cho nghiên cứu này. Hiệu quả và tính linh hoạt đã được chứng minh của nó làm cho nó trở thành nền tảng phù hợp cho hệ thống đang được phát triển.

1. INTRODUCTION

Fire incidents are one of the leading causes of property damage, loss of life, and sometimes even hinder the development of an area. However, in certain cases, detecting smoke and fire alarms can become challenging, especially in hard-to-reach locations or in dusty and hazy environments. Vietnam is one of the countries with the highest number of fire incidents in the Southeast Asia region. In the article titled "The Haunting Fires of 2021," journalist Van Ngan writes, "Throughout the country, thousands of small and large fires occurred, some of which had extremely serious consequences" (Van Ngan, 2021) [1]. According to statistics from the Ministry of Public Security (Ministry of Public Security, 2023), from 2014 to 2019, there were approximately 11,000 fire incidents, including over 1,000 major fires, causing severe damage to property and human lives. However, the annual report on fire prevention and firefighting by the Ministry of Public Security reveals that only around 20% of buildings in Vietnam are equipped with modern fire alarm and firefighting systems. This underscores the necessity of researching and developing smoke detection and fire alarm solutions. On September 13, 2023, the Ministry of Public Security issued a directive to the Firefighting and Rescue Police Department and local police departments to enhance the effectiveness of state management in fire prevention, firefighting, and rescue operations. The directive aims to address fire and explosion situations and minimize casualties and property damage, particularly in densely populated residential areas like mini apartments and high-density rental service businesses. Detecting smoke and fire alarms is crucial in Vietnam to minimize fire-related losses and ensure the safety of the population. Currently, there is a growing demand for modern smoke detection and fire alarm systems, especially in industrial zones and densely populated residential areas. Detecting smoke and fire alarms is a critical issue in ensuring community safety.

The World Health Organization (WHO, 2018) [2] reported that in 2018, approximately 180,000 people died globally due to fires, and this number continues to rise over time. In 2019, in Europe, wildfires in Portugal and Spain devastated thousands of hectares of forests, causing significant environmental and economic damage. Additionally, the presence of smoke from these fires also impacted air quality and the health of the population. The article by Pham and colleagues proposes a method to assess fire risk and design firefighting systems for universities. This method utilizes multi-state assessment techniques and evaluation criteria such as independence, dispersion, functionality, and efficiency. The results indicate that this approach can enhance the fire safety level for tall building structures. In the past, fire alarms were often given through the use of bells or sirens to alert everyone in a potentially dangerous area. However, this method may not guarantee effectiveness in certain situations, especially in large buildings or when there are many people present. Additionally, manually activating fire alarms can also pose challenges, particularly when people are not familiar with how to use them or when there are not enough alarms distributed throughout the entire area. Furthermore, the use of manual fire alarms can lead to errors as users may not recognize the hazardous situation or may not use the alarms correctly. This highlights the importance of modernizing and improving fire detection and alarm systems to ensure quicker and more accurate responses in case of fire emergencies. The article by Rongbin Xu investigates the impact of the large wildfires that occurred in Australia during the 2019-2020 summer on human health and visibility. The research findings indicate that smoke from the wildfires had a significant adverse effect on the health of people in cities located hundreds of kilometers away. Health risks increased, especially for individuals with pre-existing respiratory conditions. Additionally, the smoke reduced visibility in major cities, affecting daily life and transportation. In the article of Khan Muhammad and colleagues, the authors propose a cost-effective CNN architecture for fire detection in surveillance videos. The model draws inspiration from the GoogleNet architecture and is specifically fine-tuned to focus on computational complexity and detection accuracy. Through experiments, it has been demonstrated that the proposed architecture outperforms existing fire detection methods based on manual features as well as those based on the AlexNet architecture. The article of Panagiotis Barmpoutis and colleagues (Panagiotis, 2014) [3] introduces a novel method for smoke detection in videos. This method aims to distinguish smoke from moving objects by applying spatio-temporal analysis, smoke motion modeling, and dynamic texture recognition. It initially identifies candidate smoke regions in a frame using background subtraction and color analysis based on the HSV model. Subsequently, a spatio-temporal smoke model, including spatial energy analysis and spatio-temporal energy analysis, is applied in candidate regions. Gradient orientation histograms and optical flow (Hoghofs) are computed to capture appearance and motion information, while dynamic texture recognition is applied within each candidate region using linear dynamic systems and a bag-of-features approach. Finally, a combined dynamic

saliency score is used to determine the presence of smoke in each candidate image region. Experimental results presented in the article show the significant potential of the proposed method.

In the article titled Fire Safety Search (2020) [4], the authors introduce fundamental elements, actions, and best practices related to effective early warning systems. It supports the development and implementation of early warning fire detection systems that prioritize human-centric, timely, and understandable alerts for at-risk individuals, including guidance on how to act upon receiving an alert. In the article by Penghui Dong, a fuzzy comprehensive evaluation and Bayesian network-based method is proposed for building fire risk assessment. The advantages of Bayesian networks in handling uncertain problems have been used to model and analyze fuzzy issues that can be well addressed in practical scenarios. The Bayesian network model is analyzed using the example of a high-rise residential building fire accident on Yuyao Road in Shanghai. Prior probabilities come from the statistical history of fire incidents in high-rise buildings in recent years, and fuzzy probabilities are provided by experts. In the article by Myoung-Young Choi and Sunghae Jun (Choi and Jun, 2020) [5], the authors propose to build a fire risk model using statistical machine learning and optimized risk indexing. Data related to fire risk factors, including explanatory variables (X) influencing the occurrence of fire incidents and responses (Y) indicating the frequency of fire incidents, were collected. The research shows that the proposed model can predict fire risk with high accuracy and could be useful for fire safety management in high-rise buildings. The article by Souad Kamel and colleagues in South Korea focuses on developing an efficient system for early fire management to prevent material and human losses. This system utilizes Internet of Things (IoT) technology, making it feasible with the advancement of IoT. It combines low-cost IoT sensors to collect real-time data (such as temperature, the number of people at the scene of a fire, etc.) and presents all sensor readings on a single web-based dashboard. When values collected exceed specific thresholds, the system sends alerts to the building manager, allowing them to notify authorities or dispatch firefighters in real-time. A crucial aspect of this system is its ability to monitor the number of people at the scene of a fire, simplifying evacuation processes and enabling civil defense agencies to manage resources effectively. The system has been successfully tested in various scenarios within an educational building (Al-Faisaliah Women's Campus, Jeddah University, Saudi Arabia). In the article by Bá Tuấn and Văn Ngọc (Bá Tuấn - Văn Ngọc, 2021) [6] in Vietnam, the authors propose a fire risk prediction model for high-rise buildings using deep learning techniques. To build the model, they utilized a dataset of fire incidents in high-rise buildings collected from various sources. The experimental results demonstrate that the proposed model has the potential to accurately predict fire risk and could be valuable for fire safety management in high-rise buildings in Vietnam. An IoT-based system is applied for monitoring and automatic fire warnings: the system includes a central fire alarm cabinet, fire alarm devices, control devices, IoT Wi-Fi devices, and remote monitoring software for users. The system uses

LPWAN (Low Power Wide Area Network) IoT technology to connect wireless sensors for real-time data transmission to the cloud, helping to pinpoint the location of a fire outbreak (Science and Development, 2021) [7].

Continuous learning is currently a crucial approach in the field of artificial intelligence, particularly when dealing with data in constantly changing environments. In this project, artificial intelligence will be used in conjunction with a sliding window technique along with the Smoke Detection Dataset collected by Stefan Blattmann in the "Real-time Smoke Detection with AI-based Sensor Fusion" project and updated by Deep Contractor (comprising 62,630 rows with 15 data fields, measured by IoT sensors, including parameters such as temperature, humidity, brightness, and various other metrics) to develop models for classifying smoke and non-smoke from sensor signals collected by IoT devices.

To contribute to addressing the issue of fire and explosions, utilizing the collected and analyzed database to develop the main algorithms for a smoke detection and fire alerting system is a promising approach. By applying artificial intelligence and machine learning technologies using the mined database, organizations can have additional tools, research insights, and deployment results for addressing this problem through artificial intelligence systems. The research project aims to find a suitable method that can serve as the core algorithm for developing a sensor-based recognition system. The purpose of this research is to leverage artificial intelligence systems to provide the most accurate fire alert predictions, benefiting the social life in Vietnam. This project holds the potential to significantly contribute to minimizing the damage caused by fires and explosions, especially in emergency situations.

2. MATERIALS AND METHODS

During the data search process for the topic, numerous datasets were found (approximately 232 datasets related to smoke detection). However, there are 8 datasets specifically focused on smoke recognition for fire detection and alerting, which possess comprehensive parameters and the highest level of availability.

The "Smoke Detection using Classification Models" dataset by Sandesh Singh comprises two folders containing different types of data. The first folder contains 3423 images related to smoke and non-smoke, while the second folder contains 316 videos related to the images in the first folder. This dataset is used to build classification models for smoke and non-smoke based on machine learning algorithms. "Firesense by Chris Gorgolewski" is a dataset that consists of 14 data fields and over 200,000 rows. It provides information about forest fires, climate factors, geography, and the environment. The purpose of this dataset is to assist in predicting and managing forest fires more effectively.

The "Sensor-Fusion Smoke Detection Classification" dataset by Gaurav Dutta was updated and published in 2020. This dataset contains 1074 rows with 8 data fields. The data fields include parameters such as acceleration, angular velocity, and altitude measured using sensors. The data was collected during fire simulation to simulate fire alarm situations and smoke detection. This dataset is used to train classification models for smoke and non-smoke based on sensor signals. The "Smoke Detection Dataset" by Deep Contractor was collected by

Stefan Blattmann in the project "Real-time Smoke Detection with AI-based Sensor Fusion" and updated by Deep Contractor (Contractor, 2020). This dataset comprises 62,630 rows with 15 data fields measured using IoT sensors, including parameters such as temperature, humidity, brightness, and several others. The purpose of this dataset is to develop classification models for smoke and non-smoke based on sensor signals. This data is used for smoke detection and triggering fire alarms.

The "Fire and Smoke dataset" was collected by DataCluster Labs (DataCluster, 2023) [8]. It contains over 7,000 images of fire and smoke. Each image in this dataset has been manually reviewed and verified by computer vision experts at DataCluster Labs. This dataset is large, high-resolution (98% of images are HD or higher), and collected from over 400 urban and rural areas. It can be used for early fire and smoke detection, smart camera systems, fire and smoke alarm systems, and more. The "WildFire-Smoke-Dataset-Tensorflow" was created by Aluru V N M Hemateja, a student at Vellore Institute of Technology in Chennai, Tamil Nadu, India (Hemateja, 2021) [9]. This dataset contains images of smoke from wildfires collected from various sources. It consists of a total of 1,500 images with a size of 256x256 pixels and is divided into two folders, "smoke" and "non-smoke." The images in the "smoke" folder depict smoke from wildfires, while the images in the "non-smoke" folder do not have any relation to wildfires. The article "Forest Fire prediction using Machine Learning" on the Analytics Vidhya website was written by Aman Preet Gulati and published in the paper "Forest Fire prediction using Machine Learning." (Gulati, 2021) [10]. In this article, it is mentioned that the dataset contains 36,011 rows and 15 columns. The article introduces the use of Machine Learning to predict the likelihood of forest fires based on certain attributes. It also discusses the importance of predicting forest fires and the benefits of using Machine Learning for this purpose.

The datasets mentioned above are relatively good, but they were not used for classification purposes and some of them are outdated, making them unsuitable for the current research. To be usable for this research, the dataset must be numerical, well-classified, and have multiple fields to yield the most objective results. The Smoke Detection Dataset is the only one that meets these requirements, which is why it was chosen for this project's system. The dataset used in this study must fulfill specific criteria, including being numerical, well-classified, and having multiple fields to ensure the most objective results. Among the available options, the Smoke Detection Dataset is the most suitable, making it the primary dataset for this research's smoke detection and recognition system. The dataset's distinctive features are sourced from the aforementioned data and have been verified by domain experts in this field. Each feature will have the following metrics:

UTC (Coordinated Universal Time): is a time representation measured in seconds according to the UTC time standard. UTC is an international time standard used to synchronize time globally.

Temperature: Air temperature is a measurement of the degree of hotness or coldness of the air in a specific area. It is typically measured using temperature units such as

degrees Celsius (°C), degrees Fahrenheit (°F), or Kelvin (K). Air temperature has an impact on weather, climate, and the activities of humans and animals.

Humidity: Air humidity, also known as humidity or relative humidity, is a measurement of the amount of water vapor present in the air relative to the maximum amount the air could hold at the same temperature and pressure. Humidity is expressed as a percentage and can range from 0% (no water vapor in the air) to 100% (air is saturated, unable to hold any more water vapor).

TVOC: Volatile Organic Compounds (VOC) which are a group of organic substances that have a high tendency to evaporate into the air. TVOC is usually measured as a percentage. These compounds are commonly utilized to evaluate the air quality in various settings, including residential, workplace, and public environments. They are of particular concern in buildings, offices, and residential areas.

eCO2: Equivalent CO2 concentration is often calculated based on various values such as TVCO (Total Volatile Organic Compounds).

Raw H2: Raw molecular hydrogen, uncompensated (Deviation, temperature, etc.).

Raw Ethanol: Crude ethanol gas, in the form of ethanol (C2H5OH), that has not undergone processing or refinement. Ethanol is an organic compound, a type of alcohol, and a common solvent used in various industries.

Pressure: Air pressure is the force exerted on a specific area. In the case of air pressure, it is measured in units called Pascals (Pa), or other pressure units such as bar, psi (pounds per square inch), atm (atmosphere), mmHg (millimeters of mercury), etc. Air pressure plays a crucial role in various fields, such as aerospace, metrology, and industrial manufacturing processes.

PM1.0, PM2.5: Particle size < 1.0 μm: PM1.0 (Particulate Matter 1.0). Particle size between 1.0 μm and 2.5 μm: PM2.5 (Particulate Matter 2.5).

NC0.5, NC1.0, NC2.5: Particle size < 0.5 μm: NC0.5 (Number Concentration 0.5). Particle size between 0.5 μm and 1.0 μm: NC1.0 (Number Concentration 1.0). Particle size between 1.0 μm and 2.5 μm: NC2.5 (Number Concentration 2.5)

Aspect Ration: Aspect Ratio refers to the ratio between the width and height of an image or video. It is represented as a numerical value or a percentage, helping determine the shape and format of the content.

Label: These are critical and particularly important values. The label takes one of two values: "0" or "1". If the class is "smoke present," the label is 1, if the class is "no smoke," the label is 0. The predictions include a total of 17,873 data points of type 0 and 44,757 data points labeled as 1.

The experiments will be conducted through a nonlinear experimental model (Batch Learning, batch size = 100), meaning the system will have to perform 100 steps (Batch 100), with each step containing approximately 438 data points. In this project, there are two outputs: Fire Detection and No Fire Detection. Therefore, positive can be considered as Fire Detection, and negative as No Fire Detection. The indicators TP, TN, FP, FN have the following meanings respectively:

- TP (True Positive): Correctly detecting the presence of fire.

- TN (True Negative): Correctly detecting the absence of fire.

- FP (False Positive): Incorrectly detecting the presence of fire when there is none.

- FN (False Negative): Incorrectly detecting the absence of fire when it's present.

It's important to have:

These indicators are important for assessing the performance and accuracy of the model in fire detection

Incorrect Ground Truth Labels (0)	True Ground Truth Labels (1)	
FP	TP	**True positive predictions. (1)**
TN	FN	**False positive predictions. (0)**

Correct, TPR (True Positive Rate) is also known as the sensitivity or recall, and it represents the ratio of true positive predictions (correctly detecting the positive class) to the actual positive instances in the dataset. It's calculated using the formula:

$$\text{TPR} = \frac{TP}{TP+FN}$$

Correct, TNR (True Negative Rate) is also known as specificity and represents the ratio of true negative predictions (correctly detecting the negative class) to the total actual instances of the negative class in the dataset. It's calculated using the formula:

$$\text{TNR} = \frac{TN}{TN+FP}$$

After obtaining these two metrics, Balanced Accuracy is calculated using the formula:

$$Balance\ Accuracy = \frac{TPR + TNR}{2}$$

3. RESULTS AND DISCUSSION

The model used for conducting the scientific experiment has been specifically discussed in the previous section. Therefore, this section focuses on analyzing and comparing the results among the algorithms. The bar chart (both stepwise and averaged) below will provide a comprehensive evaluation of the algorithms' effectiveness. The age of the algorithm model used for comparison is set to 15 (15 data groups). To enhance objectivity, the average ratio is calculated from the results of 10 runs, i.e., the total result of 10 runs divided by 10. The dataset used in the experiment consists of two parts: one for training and the other for testing. The training dataset comprises 43,841 data rows, and the testing dataset contains 18,789 data rows. Although this dataset will vary in each experiment, the data remains unchanged (data conservation) and is reshuffled. This is achieved through the application of artificial intelligence methods, specifically algorithms combined with the sliding window approach. By utilizing charts to compare the average results of the algorithms, this method ensures the fairest representation of data authenticity during result comparison. The average experimental results of the algorithms are represented in the chart below (Figure 1).

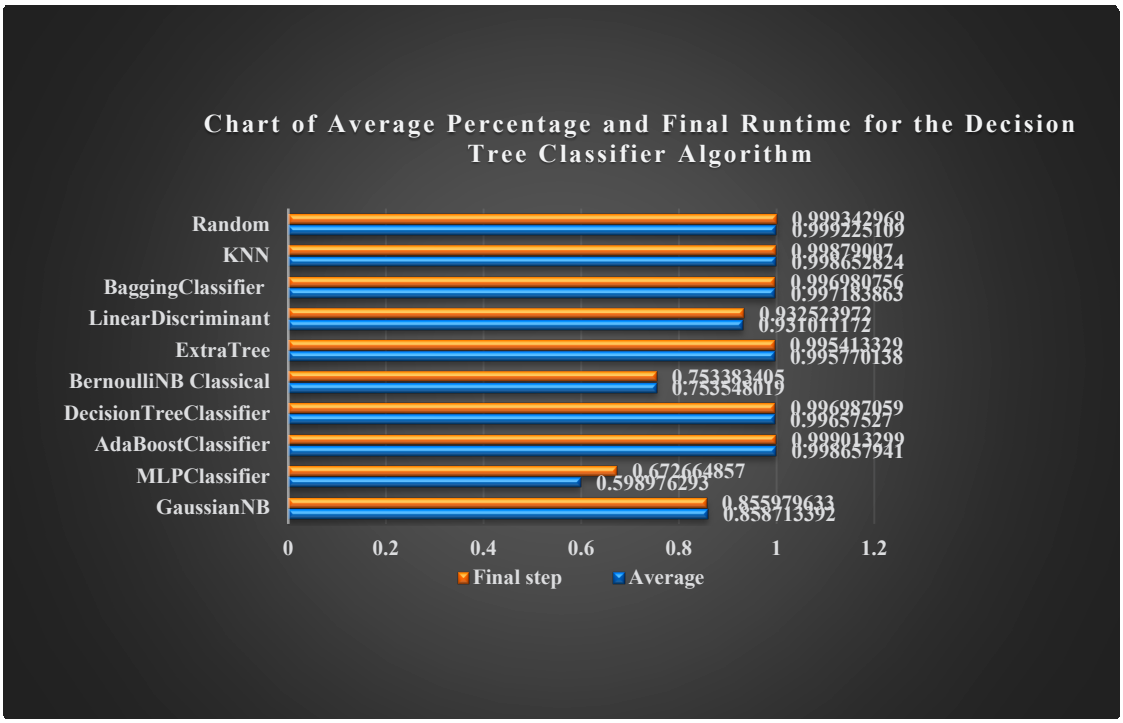

Figure 1. Average Experimental Algorithm Chart Based on Age (DecisionClassifier)

Based on the data in the chart, we can analyze the performance of the DecisionClassifier as follows:

Observing the chart, we can see that the DecisionClassifier maintains a stable performance above the 95% threshold. The average accuracy is achieved above 95% (specifically 99.70%), and the accuracy at the final step is also noteworthy, exceeding 95% (specifically 99.70%). This demonstrates the stability and ability to maintain a high-performance level of the model. This capability ensures safety and protection in situations where high sensitivity to smoke detection is needed. In addition to calculating the algorithm's average results, another approach to analyzing the effectiveness of the experimental model is based on age-wise analysis. This analysis provides a comprehensive and detailed view, helping us understand the model's capabilities for each specific age group. This approach allows us to evaluate and draw more accurate conclusions about the experimental model's performance. To illustrate this analysis, the chart below shows the results of the experimental model based on age groups (Figure 2):

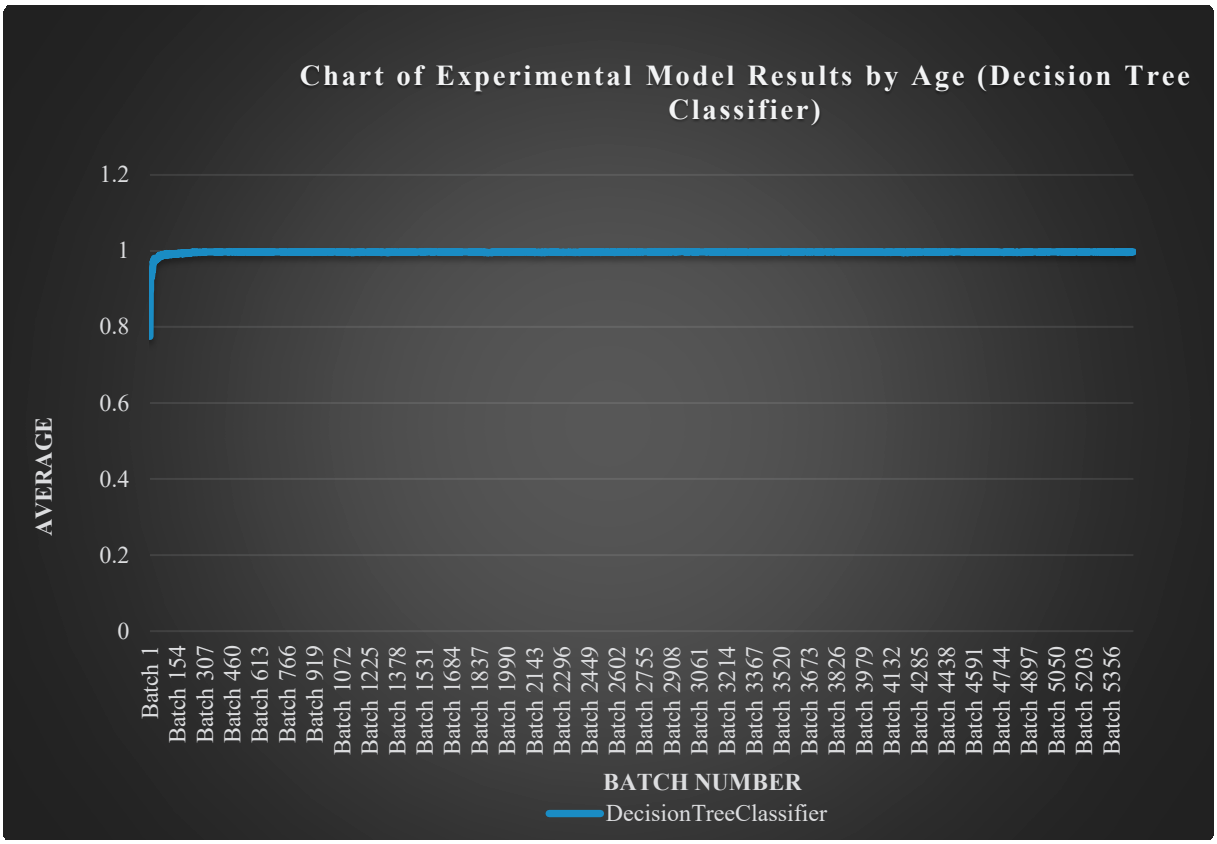

Figure 2. Experimental Model Results Chart by Age Groups (Data Groups)

Analyzing the chart, we observe that the DecisionClassifier algorithm has a relatively moderate start, achieving quite good results in the initial steps ranging from 79% to 98% over the first 84 steps, and then it starts to stabilize. Looking at the chart, we can see that the DecisionClassifier exhibits relative stability. However, it's important to note that the highest accuracy of the DecisionClassifier can reach up to 99.9%, which is a considerably high value that can be compared to many other algorithms. Specifically, during the phase from step 3505 to 3514, as shown in the chart below (Figure 3).

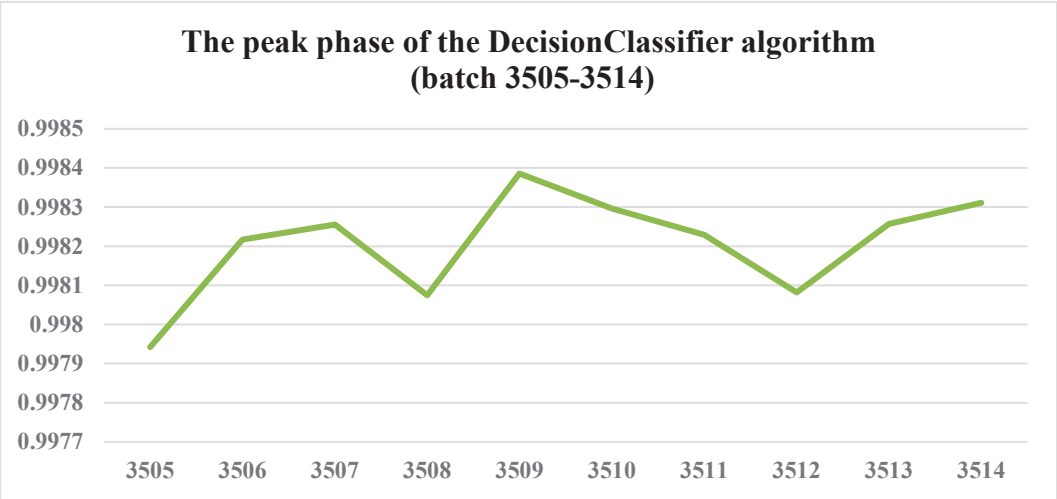

Figure 3. Highest Phase of the DecisionClassifier Algorithm (Batch 3505-3514)

The DecisionClassifier not only maintains a performance level above 99% but also excels and demonstrates greater stability compared to other algorithms. This is attributed to its ability to handle non-continuous and missing data. This capability reduces the preprocessing workload while increasing the flexibility and effectiveness of the model across various types of data. The DecisionClassifier consistently demonstrates reliability, even in the initial stages before reaching its best performance during the run, although it remains at a relatively high level. Notably, this model quickly recovers and maintains stability, a trend clearly evident in the chart below (Figure 4). The model's rapid recovery and sustained high performance following the initial phases underscore the robustness and dependability of the DecisionClassifier in data prediction and classification tasks.

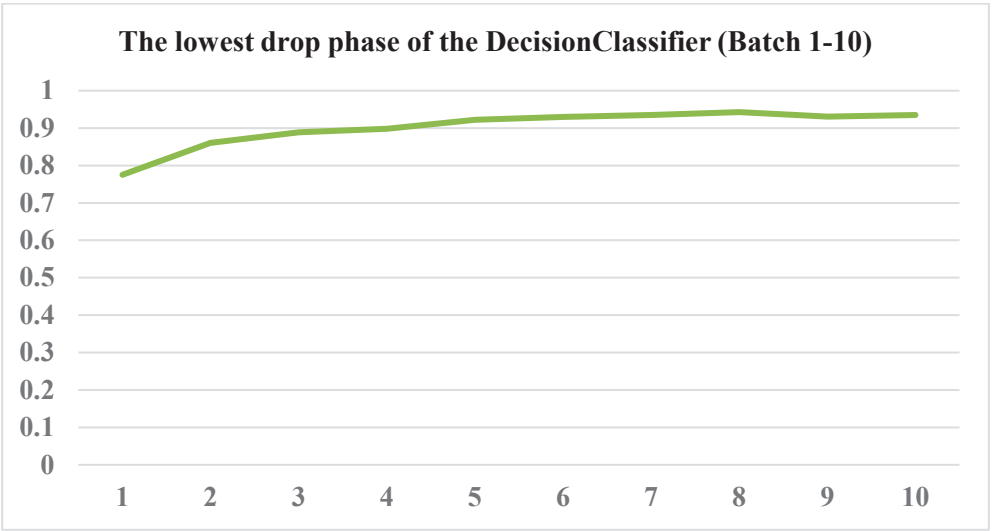

Figure 4. Lowest Drop Phase of the DecisionClassifier (Batch 1-10)

When examining the data from Batch 1 to Batch 10, we can clearly observe the variations in the achieved accuracy of the Decision Tree model. Initially, the accuracy increased from 0.793 (79.30%) to 0.880 (88.02%), creating significant fluctuations during the initial stabilization phase. However, from Batch 5 onwards, the model started to exhibit more stability and rapidly elevated the accuracy from 0.930 (93.04%) to 0.955 (95.56%) in a short time span. This demonstrates the model's ability to quickly recover and maintain stability after reaching a consistent level. In summary, within the range from Batch 1 to Batch 10, the Decision Tree model shows significant initial variations followed by maintaining a stable level and a quick recovery capability during the runtime. The model's performance continues to gradually improve and maintains a high level of accuracy after reaching a good accuracy rate. This evaluation underscores the impressive potential of the Decision Tree in prediction and classification tasks.

Installation:

The system comprises two main functionalities: algorithm installation (developer) and prediction (end user). The project will encompass functional buttons like prediction, running classic algorithms, a list of processed models, system configuration, and login. It will be implemented in a web-based environment and depicted using the use case diagram below (Figure 5).

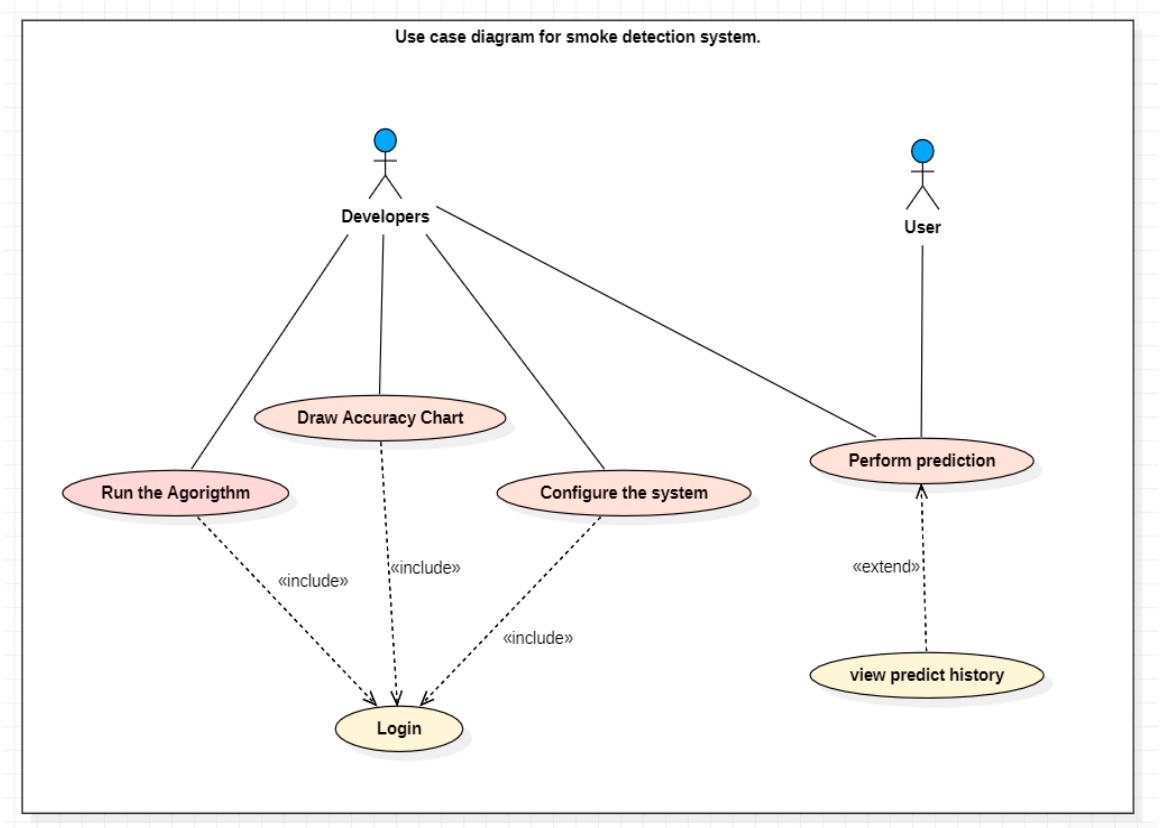

Figure 5. Use-case diagram of the system

System Configuration:

In this project, the system is packaged as a .ZIP file and compressed into a file named "BaoCaoThucTap_main". After users download and extract it, they will find a folder named "BaoCaoThucTap." Inside this folder, there are files for installation and running the program, such as "_SETUP_", requirements.txt, and RunServer.bat. The system requires a computer with a stable internet connection, minimum configuration of Windows 10, 2GB RAM, and 10GB or more of available disk space to ensure smooth and stable performance. To perform the installation, navigate to the "_Setup_" folder. The Python version required for installation is 3.9.9 (python-3.9.9-amd64.exe). Install the relevant libraries by running the CaiThuVien.bat file. The Remove.bat file can be used to delete all program data. The system can change the administrator account in 'BaoCaoThucTap_main/dataUser.csv'.

After completing the environment setup process, there will be a "Remove.bat" file, which is used to delete unnecessary data files, including those used for testing. This file should only be used in two cases: right after extraction and installation, and when users want to delete all previously run data. Run the RunServer.bat file to start the program, then open a web browser and access the address http://127.0.0.1:8000. Below is the main interface page, featuring a button to initiate the prediction process, The Start Prediction and 13 input fields to enter diagnosis data corresponding to the input parameters in the data constraints table.

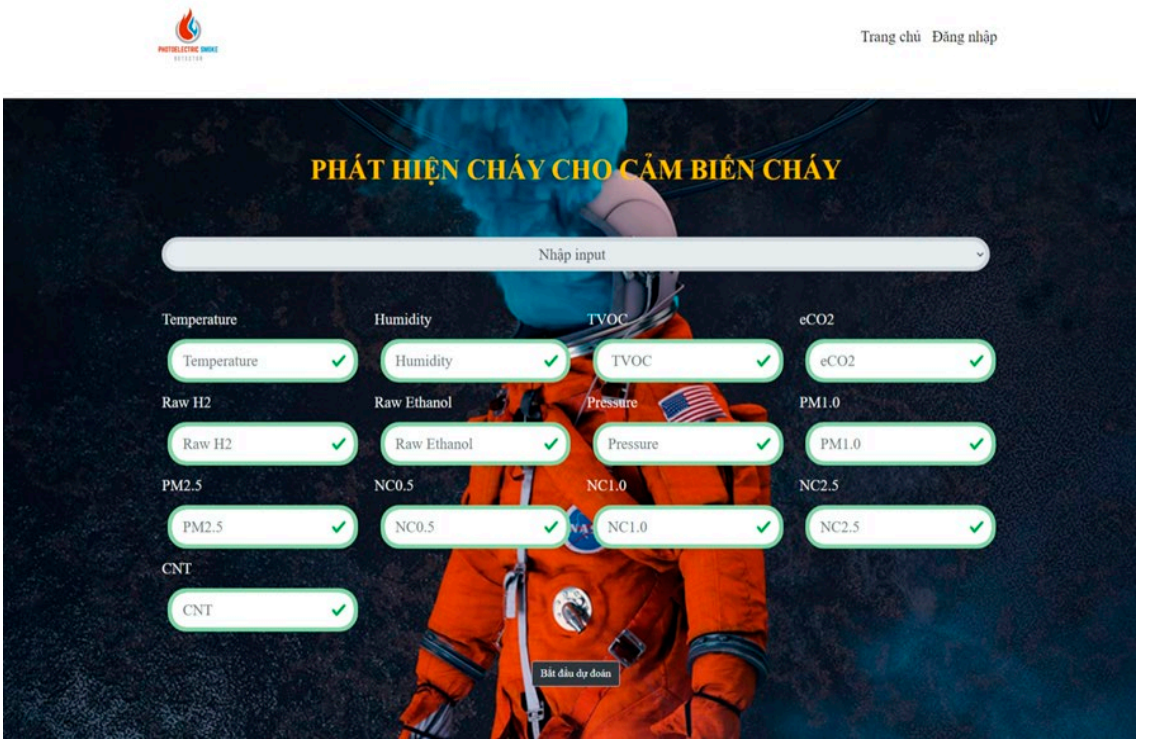

Figure 6. User Interface

4. CONCLUSION AND SUGGESTIONS

Upon concluding the research and report compilation, it can be evaluated that the achieved outcomes are reasonably comprehensive. Concerning the system, the classical DecisionClassifier algorithm has been successfully integrated into the model training and prediction process, addressing the challenge of dynamic data processing—a feat that other algorithms have not yet accomplished. Furthermore, a user-friendly graphical interface has been constructed alongside a command-line system. In terms of the report, crucial components such as data terminology, employed algorithms for training, and detailed descriptions of each chart have been meticulously covered within the content. It's important to note that this project was confined to the research phase, hence it has been developed and deployed to the extent of research. As a result, numerous avenues for expansion remain in the future. Future enhancements could encompass automated raw data processing within the system, optimizing the model training process, improving the system's interface for smoother functionality, and real-world deployment. The latter step aims to facilitate swift and precise smoke detection and fire alerting for users. Additionally, the project is also looking to expand its utility through diversification on mobile platforms, enabling users to access real-time information for recognition on-the-go via mobile devices.

REFERENCES

[1] Văn Ngân/Báo Điện Tử VOV. (2021). *Việt Nam có hơn 1.200 vụ cháy từ đầu năm đến nay.* Nguồn: Ám ảnh những vụ cháy kinh hoàng trong năm 2021 (vov.vn).

[2] World Health Organization. (2018). *Burns.* https://www.who.int/news-room/fact-sheets/detail/burns.

[3] Panagiotis Barmpoutis; Kosmas Dimitropoulos; Nikos Grammalidis (2014). Smoke detection using spatio-temporal analysis, motion modeling and dynamic texture recognition. *2014 22nd European Signal Processing Conference (EUSIPCO)*, INSPEC Accession Number: 14775489 Publisher: IEEE Conference Location: Lisbon, Portugal. https://ieeexplore.ieee.org/abstract/document/6952375.

[4] Fire Safety Search. (2020). *Very early warning fire detection-fire safety search.* https://www.firesafetysearch.com/very-early-warning-fire-detection/ (2020). Accessed on April 26, 2020.

[5] Myoung-Young Choi and Sunghae Jun. (2020). Fire Risk Assessment Models Using Statistical Machine Learning and Optimized Risk Indexing, *Appl.Sci, 10*(12), 4199. https://doi.org/10.3390/app10124199

[6] Bá Tuấn - Văn Ngọc (2021). *Thực trạng và giải pháp hạn chế nguy cơ cháy nổ trong các nhà, chung cư cao tầng. Cảnh sát phòng cháy, chữa cháy và cứu nạn, cứu hộ.* http://canhsatpccc.gov.vn/ArticlesDetail/tabid/193/cateid/1136/id/9668/language/vi-VN/Default.aspx

[7] Khoa học và Phát triển. (2021). *Giám sát và cảnh báo cháy tự động bằng công nghệ IoT.* https://congnghiepcongnghecao.com.vn/.

[8] DataCluster Labs (2023). *Fire and Smoke Dataset.* Bộ dữ liệu phát hiện sớm cháy và khói, Camera thông minh, Hệ thống báo cháy.

https://www.kaggle.com/datasets/datacluste rlabs/fire-and-smoke-dataset.

[9] Hemateja (2021). *Wild Fire-Smoke-Dataset-Tensorflow.* https://www.kaggle.com/datasets/ahemateja 19bec1025/wildfiresmokedataset

[10] Gulati, A.P. (2021). *Forest Fire Prediction Using Machine Learning.* https://www.analyticsvidhya.com/blog/202 1/10/forest-fire-prediction-using-machine-learning/.

Journal of Science and Development Economics
Nam Can Tho University

Website: jsde.nctu.edu.vn

Development experience of glass classification by Bernoulli Naive Bayes improved the continuous learning method

Nguyen Thi Cam Tu[1], Doan Hoa Minh[1], Bui Hoang Bac[2], Ngo Ho Anh Khoi[1]

[1]*Faculty of Information Technology, Nam Can Tho University*

[2]*Hanoi University of Mining and Geology*

[*]*Corresponding author: Doan Hoa Minh (email: dhminh@ctu.edu.vn)*

Received: November 10, 2023
Revised: November 30, 2023
Accepted: December 5, 2023

Keywords: *advanced machine learning, Bernoulli Naive Bayes, deep learning, glass types*

Từ khóa: *các loại kính, Bernoulli Naive Bayes, học máy nâng cao, học sâu*

ABSTRACT

Artificial intelligence is gradually emerging as a method for optimizing various tasks, offering cost-saving and highly efficient solutions. Nowadays, AI is used as a general term for diverse tasks performed by computers. Fields like machine learning, deep learning, and data science, among others within this scope, are considered part of AI as long as they exhibit the characteristics of artificial intelligence. AI is particularly valuable in predictive analysis, specifically in predicting datasets, and is applied to classification problems. The application of artificial intelligence in solving the glass classification problem aims to categorize and recycle different types of glass. The sliding window method is employed for this classification task as it is the most suitable approach. By classifying glass, this approach contributes to the recycling and reuse of industrial glass, reducing glass waste for the benefit of humanity and limiting environmental pollution.

TÓM TẮT

Trí tuệ nhân tạo đang dần nổi lên như một phương pháp tối ưu hóa các nhiệm vụ khác nhau, đưa ra các giải pháp tiết kiệm chi phí và hiệu quả cao. Ngày nay, AI được sử dụng như một thuật ngữ chung cho các nhiệm vụ đa dạng được thực hiện bởi máy tính. Các lĩnh vực như học máy, học sâu và khoa học dữ liệu, cùng những lĩnh vực khác trong phạm vi này, được coi là một phần của AI miễn là chúng thể hiện các đặc điểm của trí tuệ nhân tạo. AI đặc biệt có giá trị trong phân tích dự đoán, cụ thể là dự đoán các tập dữ liệu và được áp dụng cho các bài toán phân loại.

Việc ứng dụng trí tuệ nhân tạo vào giải bài toán phân loại kính nhằm mục đích phân loại và tái chế các loại kính khác nhau. Phương pháp cửa sổ trượt được sử dụng cho nhiệm vụ phân loại này vì đây là phương pháp phù hợp nhất. Bằng cách phân loại kính, phương pháp này góp phần tái chế và tái sử dụng kính công nghiệp, giảm thiểu rác thải thủy tinh vì lợi ích của nhân loại và hạn chế ô nhiễm môi trường.

1. INTRODUCTION

With the rapid development of industrialization and modernization, human needs have also increased accordingly. In fields such as construction and aesthetics, glass plays a significant role in fulfilling essential human requirements. Glass appears in various types of structures, ranging from religious edifices like churches to office skyscrapers and the private living spaces of families. The application of glass is diverse and prevalent to the extent that in any modern house, we can observe this material. In modern life, glass is extensively used due to its beauty and quality. Most modern-style houses utilize glass for decoration and construction. The advantages of glass lie in its user-friendliness, ease of cleaning, lightweight nature, and its ability to transmit light. When we fully comprehend the functional aspects of glass and use it appropriately, it enhances its aesthetic value. Presently, the market offers a wide array of glass types, but glass recycling still remains less common. Recognizing the need for classifying and recycling various types of glass is an intriguing subject that contributes to the development of our domestic industry. Hence, the topic "Advancing Machine Learning for the Identification of Glass Types" has been chosen for the reason that if successfully applied and widely developed, it will bring numerous benefits to the glass industry as well as other domains like construction and beauty enhancement. Moreover, it will assist in reducing industrial glass waste, yielding advantages for both humanity and the environment.

Some international articles on the significance and benefits of glass recycling are discussed in, the author emphasizes the best options for large and small-scale recycling, reusing, and repurposing of public glass, highlighting the importance of glass and glass recycling for both human beings and the environment. In Rinkesh et al. (2023) [1], the author elaborates on the glass recycling process and its benefits, as it can be transformed into various products for daily human use. It is an amorphous solid that can have different semiconductor components, but most importantly, it is made from molten silica along with limestone and soda ash,... Another article titled 'Recycling glass is one of the many ways we can help reduce pollution and waste,' discusses how glass recycling contributes to environmental well-being, specifically how recycling glass is one of the ways to help minimize pollution and waste. Every day, tons of waste are discarded, and glass constitutes a significant portion. Instead of allowing landfills to accumulate hazardous glass items that threaten safety and the environment, we can

reuse them, emphasizing the benefits of glass recycling for the environment. In Tạp Chí Môi Trường (2023) [2], the author explains the reasons for recycling waste, specifically glass, and its benefits, the author also addresses the issues of the harmful nature of glass, whether broken glass items should be recycled, and the specific methods of recycling broken glass items in the article 'Should broken glass items be recycled?'. An environmental journal article also highlights the reasons for classifying and recycling glass waste in the specific article 'Why glass waste should be classified and recycled' [2].

In Okafor (2023) [3] has outlined the ways of recycling glass that anyone can do on their own. Specifically, the author has identified the types of glass that can be recycled and those that cannot, along with the methods to recycle them. Furthermore, the author has also highlighted the environmental impacts of improper glass disposal, stating that it takes a million years for glass to completely decompose. Despite the numerous benefits of glass recycling, not all types of glass are recyclable. Over the course of many years of decomposition, glass breaks down into small fragments of super-small and nano sizes. Scientists are concerned about the potential hazards of nano glass particles corroding the environment. Nano glass pollution poses a greater environmental threat than pollution from larger glass particles. Glass at the nano scale becomes the end of the food chain, ingested by marine life. Additives and other components used in glass production can harm both terrestrial and aquatic animals when consumed. Additionally, plant roots can absorb these super-small glass elements. Environmental glass pollution can lead to soil degradation, loss of animal habitats, and water contamination [3]. In Cleanipedia (2023) [4], there are 27 ways to recycle old glass bottles into useful home decorations are pointed out, such as recycling glass bottles into flower vases, repurposing glass bottles as water jugs, upcycling glass jars into chandeliers, reusing glass containers as storage items, utilizing glass bottles as plant pots, creating mesmerizing twinkling effects using glass bottles, making alcohol lamps, and crafting oil lamps by recycling glass bottles [4]. Thu Ngân (2022) [5] presented various methods of recycling glass bottles, such as transforming glass bottles into bird food troughs, repurposing glass bottles as wind chimes, simple decoration ideas for glass jars like turning them into oil lamps, repurposing glass bottles as decorative items for weddings [5].

Currently, there are numerous solutions worldwide addressing the issue of glass recycling. In Experts (2022) [6], five pieces of advice are provided for recycling glass, focusing on specific aspects. For instance, shattered glass can indeed be recycled, but it might not return to its original state. In other words, recycling a broken bottle may not result in the glass being remanufactured into a new bottle. Instead, the glass can still find utility as an additive in glass fibers or tiles. However, it's crucial to maintain the integrity of recycled glass as much as possible. Regarding glass handling, not all glass is the same. Glass utilized for windows, mirrors, and similar items undergoes chemical treatment and thus possesses a distinct melting point compared to, for example, glass bottles. Consequently, it's generally advised not to mix non-container glass with container glass during recycling.

Additionally, it's safer to avoid processing shattered glass whenever possible. Many recycling facilities require glass to be cleaned before recycling. If the glass contains residues, such as sugar, for instance, it can become sticky and potentially attract pests. This is also true for other glass containers used for food and beverages. In Momentum Recycling (2023) [7] the steps for glass recycling and the resulting products are highlighted for practical applications. These steps include glass recycling processes, collection, sorting stations, glass breaking, trommel, steam layer drying, filtering and cleaning, crushing, secondary screen size classification, and final product output. The author also presents various products applied after recycling glass, such as glass container production, glass fiber manufacturing, abrasive material, fluxing agent in ceramics and bricks, filler in paint and plastics, glass bead for reflection, adsorbent material, and cation exchange [7].

In Vietnam, the recycling of glass to minimize environmental harm is increasingly receiving positive attention, as highlighted in numerous articles by khoahoc.tv. These articles address questions such as the feasibility of glass recycling and the specific methods involved. Recycling used glass bottles, for instance, proves to be an effective means of safeguarding our planet. With each ton of recycled glass, humanity conserves a significant amount of raw materials necessary for producing new glass, including 590 kg of sand, 186 kg of Sodium Carbonate powder, and 173 kg of limestone. The manufacturing of new glass also consumes substantial energy and contributes to industrial pollution, thus exacerbating the greenhouse effect. This is due to the requirement of heating sand and other substances to temperatures exceeding 1400°C to create glass. Conversely, recycling glass consumes less than 40% of the energy required for producing new glass. The glass is collected and sorted by color—basic hues of white, green, and amber—then cleaned to remove contaminants. During the sorting process, paper labels adhered to glass containers are removed, along with all non-glass elements such as plastic caps, metal lids, and types of glass that cannot be recycled. After the initial sorting, the glass is shattered using crushing machines, followed by passing through machines that separate metal, plastic, and paper. The glass is then ground into small fragments called cullet. The purpose of grinding the glass is to eliminate sharp edges that pose hazards. Subsequently, the fragments are sifted to filter out larger pieces for further grinding, resulting in glass particles of the desired size. Cullet refers to the finely ground glass, free from non-glass materials, which is ready to be fed into mixing machines for crafting new glass (Khoahoctv, 2023) [8].

The following, the author will employ the sliding window approach to address this issue by using Machine learning, through the Glass Types database, across the sections: the theoretical foundation of the issue, the implementation method, and the experimental results of the problem, before moving to the conclusion.

2. MATERIALS AND METHODS

Based on glass categorization, it aids in the recycling and reutilization of various types of industrial glass, contributing to reducing glass waste for humans and limiting environmental pollution. The problem is addressed using the sliding window method as it is the most suitable

approach for this problem. The existing database remains static over time, trained using classical methods (performed only once, requiring retraining from scratch with new data). However, in modern reality, data environments change over time, necessitating continuous real-time training and periodic updating of predictive models. Hence, data learning must be carried out within an ever-changing data environment, indicating the practical application of continuous learning in non-stable environments. This problem aligns with scenarios involving evolution within an unstable environment. To address this, the term "Concept drift" has been widely adopted. The concept of drift forms the basis for slow, continuous change and the "forgetfulness" of past situations. Nevertheless, the challenges in an unstable environment stem from the fact that they are contingent on various factors. Developments may sometimes occur rapidly, at other times slowly; sometimes forgetting occurs, and there are even instances where knowledge resurfaces after it has disappeared. It should also be noted that these approaches are not sufficient standards for an "incremental approach." The document lists three different methods proposing solutions to this issue: The Sliding Windows-based approach, considering the evolution of concepts in an environment of recent non-stationary training data, determined by a defined time window (according to a time scale or a data quantity). This approach can reclassify the "group" type (on data selected by the temporary window), or update the model if online learning methods allow. In this case, the "forgetting" (as mentioned above) is automatically managed by this learning method. This type of approach usually consists of 3 steps: 1) detecting concept changes by using statistical tests on different windows; 2) if an observed change exists, select representative and recent data to adjust the models; 3) update the models. The window size is predetermined by the user. The main point of these methods is to determine the window size. Most methods use a fixed-size window configured for each real-world problem. This issue will rely on the aforementioned characteristics and based on the history of research on the methods, the research will employ the “sliding window” method for the upcoming task, specifically using the “sliding window” method on the Bernoulli Naive Bayes algorithm.

During the process of searching for data for the topic, a multitude of datasets were found. However, these datasets lack complete specifications and the highest level of availability. These datasets include: "multi-Classification of Glass Types" by Jay "Glass Types Classification Tensorflow ResMLP” (GeeksforGeeks, 2021) [9] and "Glass Types Predict with SMOTE". The aforementioned datasets are all related to glass detection or classification; however, only one dataset meets the requirements of the topic. Other datasets have various issues that render them unusable for this particular project. For a dataset to be usable in this project, it must be numerical data, exhibit specific class divisions, contain multiple attributes to yield objective outcomes, and have recently updated and relevant data. The "Glass Types" dataset is the only one that fulfills these criteria, utilized for solving this classification problem making it the chosen dataset which is provided by the author Zahra Arabi, a member of the Kaggle website. The most recent update

to this dataset was in August 2022, based on the Vietnam local time. The creation of the "Glass Types" dataset aims to address the problem of glass recycling classification, facilitating the classification and reuse of various glass types. The most recent update to the dataset was in August 2022, according to the Vietnam local time.

The raw data must undergo preprocessing before being applied to the software's training process. From the raw data, unnecessary information such as serial numbers and IDs will be removed, as they are not essential during program execution. The dataset must be saved in a ".csv" file format. Each data entry is allowed only in a single cell, where parameters and labels must be separated by a comma "," as stipulated by the system's regulations. According to the system's guidelines, the data parameters must be placed at the front, and the label should be placed at the end. After all the data standardization processes, we will obtain a normalized dataset consisting of 9 features, totaling 215 data entries, including parameters in the following format:

<Label Predictions> 1: <RI> (Refractive Index): In reality, each RI value corresponds to a different size, while factors such as material, color, style, etc. will be completely consistent. The range of RI values can go from 1 to infinity. The size of the glass being predicted is a mandatory input, and the minimum and maximum sizes in the dataset are 1.5115 and 1.53393, respectively.

2: <Na> (Sodium): An elemental component of glass, sodium is a chemical element as well as one of the constituents of glass. The minimum and maximum sodium values in the dataset are 10.73 and 17.38.

3: <Mg> (Magnesium): An elemental component of glass. Magnesium, the 8th most abundant element in Earth's crust, is an alkali metal. The minimum and maximum magnesium values in the dataset are 0 and 4.49.

4: <Al> (Aluminum): An elemental component of glass. Aluminum, the most abundant metal in Earth's crust, constitutes about 17% of Earth's solid outer layer. The minimum and maximum aluminum values in the dataset are 0.29 and 3.5.

5: <Si> (Silicon): An elemental component of glass. Silicon is a very hard element with a dark gray - metallic blue shine. The minimum and maximum silicon values in the dataset are 69.81 and 75.41.

6: <K> (Potassium): An elemental component of glass. Potassium is a soft alkali metal with a silver-white color that easily oxidizes in the air. The minimum and maximum potassium values in the dataset are 0 and 6.21.

7: <Ca> (Calcium): An elemental component of glass. The minimum and maximum calcium values in the dataset are 5.43 and 16.19.

8: <Ba> (Barium): An elemental component of glass. The minimum and maximum barium values in the dataset are 0 and 3.15.

These indices are entirely derived from the above data source and have been validated by experts in the field. Each feature will have values as follows:

− RI (Refractive Index): In reality, each RI value corresponds to a different size, while factors such as material, color, style, etc. will be completely consistent. The range of RI values can go from 1 to infinity. The size of the glass being predicted is a mandatory input, and the minimum and maximum sizes in the dataset are 1.5115 and 1.53393, respectively.

− Na (Sodium): An elemental component of glass, sodium is a chemical element as well as one of the constituents of glass, and the minimum and maximum sodium values in the dataset are 10.73 and 17.38.

− Mg (Magnesium): An elemental component of glass, and the minimum and maximum magnesium values in the dataset are 0 and 4.49.

− Al (Aluminum): An elemental component of glass, and the minimum and maximum aluminum values in the dataset are 0.29 and 3.5.

− Si (Silicon): An elemental component of glass, and the minimum and maximum silicon values in the dataset are 69.81 and 75.41.

− K (Potassium): An elemental component of glass, and the minimum and maximum potassium values in the dataset are 0 and 6.21.

− Ca (Calcium): An elemental component of glass, and the minimum and maximum calcium values in the dataset are 5.43 and 16.19.

− Ba (Barium): An elemental component of glass, and the minimum and maximum barium values in the dataset are 0 and 3.15.

− Fe: an element constituting glass. Iron is a useful element on Earth, forming the outer and inner layers of the Earth's core, and the smallest and largest values of aluminum in the dataset are 0 and 0.51, respectively.

3. RESULTS AND DISCUSSION

The dataset is provisional due to the emergence of various new types of glass, leading to possible changes in the dataset's classification in the future, indicating its instability. Consequently, increasing the dataset's dimensions cannot be applied in classical machine learning algorithms within a static environment, as traditional methods do not allow dimension expansion, necessitating the use of progressive algorithms capable of handling data in dynamic environments. Upon downloading the dataset, the values will not be entirely appropriate, thus requiring normalization. Following this, data will be divided to create proportions for training and testing, with a ratio of seven parts (70%) to three parts (30%). After segmenting the data, it will be saved to the machine, then the proportions will be allocated for the execution of the algorithm. This will be done by selecting a batch size of 70% * n or 70% * n/2, with a training/test ratio of 70%, where n represents the total number of data. All achieved results are based on Balanced Accuracy, a metric used to assess the performance of binary classifiers. It proves particularly useful when dealing with imbalanced classes, meaning one of the two classes is far more frequent than the other. The utilization of Balanced Accuracy is significantly more intricate than conventional Accuracy. Conventional Accuracy simply calculates the percentage ratio of a dataset based on the total available data, which initially functions stably. However, when the data is greatly skewed, the accuracy's reliability diminishes. To address this issue, the formula for Balanced Accuracy can be employed to calculate percentages in the most authentic and optimal manner.

By applying artificial intelligence, specifically algorithms such as Bernoulli Naive Bayes in combination with the sliding window method, this approach achieves the most fairness in terms of data veracity when comparing results from different algorithms. The experimental outcomes of the algorithms are averaged and presented in the chart in Figure 1.

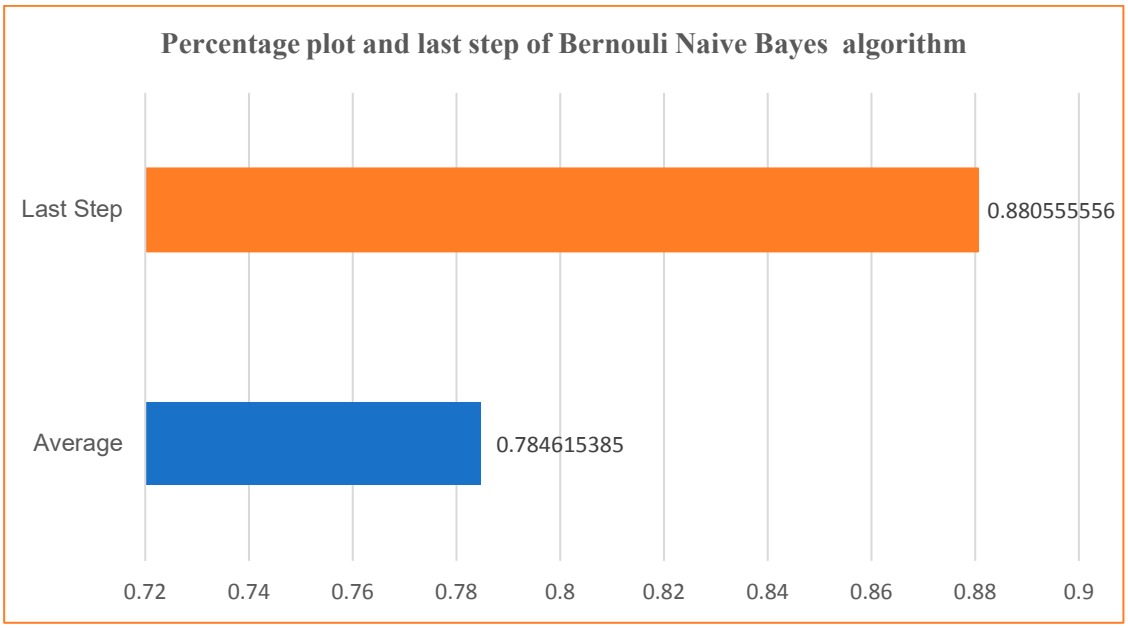

Figure 1. The chart of average percentages of experimental algorithm performance by age (Bernoulli Naive Bayes)

Looking at the data on the chart, we can analyze the average ratio, and the final step has a fairly good ratio of Bernoulli Naive Bayes as follows:

The performance of Bernoulli Naive Bayes is very stable and consistently reaches above 70%. The data demonstrates that Bernoulli Naive Bayes achieves a steady and remarkable performance with an average accuracy of over 70% (precisely 78.61%), and the accuracy in the final stage also surpasses 80% (precisely 88.06%). The performance of the final step and the average, with a relatively narrow gap of 9.45%, indicates a fairly stable algorithmic performance. This is an important advantage, highlighting that Bernoulli Naive Bayes has the ability to provide accurate predictions in various scenarios, making it applicable to this predictive task.

In addition to calculating the algorithm's average results, another approach such as analyzing the experimental model results by age group provides a more comprehensive and detailed perspective. This helps us visually assess and reach the most accurate conclusions regarding the experimental model's results by age group, represented in the chart in Figure 2.

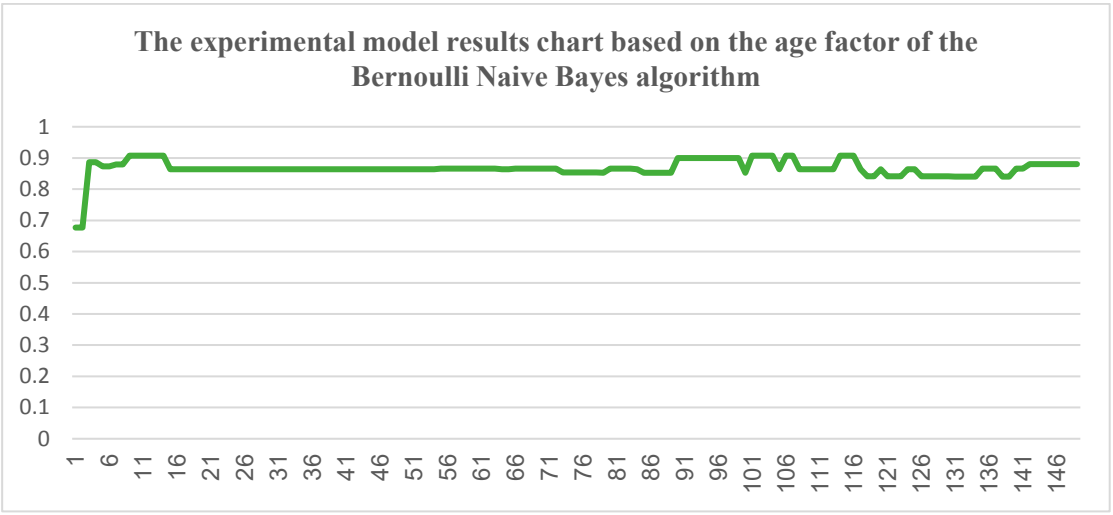

Figure 2. Chart of Experimental Model Results by Age (Bernoulli Naive Bayes)

Examining the chart, we can observe that the Bernoulli Naive Bayes algorithm starts with a relatively stable point at around 67.69%, and then experiences a sharp increase in the subsequent steps, reaching 90.04% from step 2 to step 6. Following this, the algorithm maintains a consistently stable and fairly uniform performance in steps 7 to 16, specifically at 86.40%. From step 17 to step 71, the algorithm demonstrates stability, maintaining a consistently high performance represented by a horizontal line on the graph, specifically at 86.40%. Overall, the Bernoulli Naive Bayes algorithm exhibits a stable and quite high performance across the entire chart, with its peak performance occurring in steps 111 to 121, as shown in the graph in Figure 3.

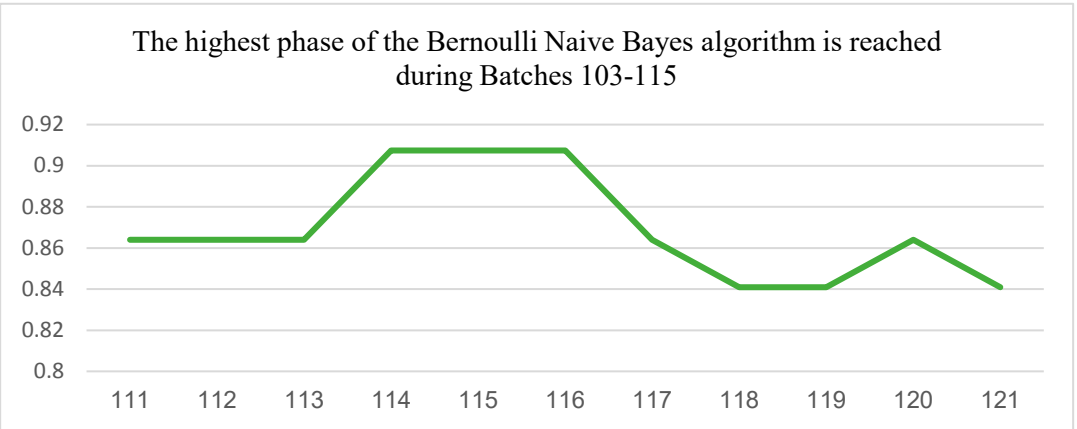

Figure 3. Experimental Model Results by Age (Bernoulli Naive Bayes)

Examining the chart, we clearly observe a consistent upward trend of Bernoulli Naive Bayes's performance across different age groups. The accuracy consistently remains above 80% and nearly reaches the 95% mark. The highest points are observed at steps 114 and 115, reaching an accuracy of 90.74%. This represents a notably high rate compared to other algorithms, as Bernoulli Naive Bayes excels in handling non-continuous and missing value data. This capability significantly reduces the need for preprocessing tasks and allows the algorithm to operate effectively across various data types. Despite instances where the algorithm's results may not be optimal during certain data phases, Bernoulli Naive Bayes consistently maintains a stable and swift performance, as depicted in the chart below (Figure 4).

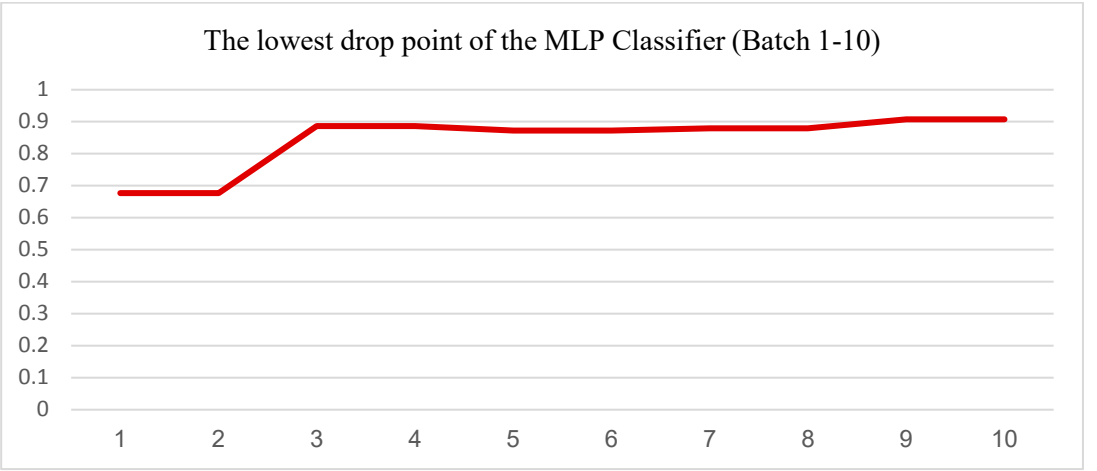

Figure 4. The lowest falling segment of the Decision Tree (Batch 795 - 815)

In steps 1 and 2, the lowest success rate of Bernoulli Naive Bayes achieved is only 67.70%. However, in step 3, this algorithm quickly returns to a stable rate, reaching 88.60% and gradually increasing in the following steps. Despite being the phase with the lowest data accuracy in Decision Trees, it is considered relatively stable as the discrepancies in numbers are not significantly large. For instance, steps 4 to 8 form an almost straight line with fairly similar ratios. Specifically, step 4 has an accuracy of 88.60%, while steps 5, 6, 7, and 8 maintain a rate of 87.29%.

In summary, the Bernoulli Naive Bayes algorithm offers numerous advantages and consistent performance, achieving an accuracy of over 90%, with the highest being 90.74%. This demonstrates that Bernoulli Naive Bayes can be a useful choice for various prediction and classification tasks. However, it's important to note that each algorithm has its own strengths and weaknesses. Selecting the appropriate algorithm also depends on the specific requirements of the task and the characteristics of the data.

Installation of the practical application:

In this topic, the system has been packaged in the form of a .ZIP file and compressed into a file named "GlassType.zip". Once users download and extract it, there will be a folder named "BaoCaoThucTap" ("InternshipReport" in English). Inside this folder, there are files

required to run the program. The system has the following computer requirements: a constant internet connection, minimum configuration of Windows 10, 2GB RAM, and 10GB hard drive space. To proceed with the installation, navigate to the "SETUP" directory. Install Python by executing the Python version 3.9.9 (Python-3.9.9-amd64.exe). Install the necessary libraries by running the script "CaiThuVien.bat".

To run the software, your computer needs to have the Node.js environment installed. Once you've opened the folder, press and hold the "Shift" key, then right-click in the empty area of the folder. Select "Open command window here" to display the command line dialogue. In the command prompt, input the command "pip install -r requirements.txt" to install the required packages for the program. After successfully installing the packages, to execute the program, you need to enter the command "py manage.py runserver" to initiate the program. The program will run on the default port "http://127.0.0.1:8000/". While the program can be deployed on a web platform, current circumstances do not permit practical web deployment. This is the default installation process for the program.

The algorithm used in this problem will be the best-performing algorithm, which is the Random Forest algorithm that was ultimately selected. The website includes functionalities such as login and logout features. The homepage functionality will comprise 7 buttons for inputting the following initial parameters: RI, Na, Mg, Al, Si, K, Ca, along with a prediction initiation button to perform diagnoses. The algorithm selection feature allows users to choose the Bernouli Naive Bayes algorithm. In the algorithm interface, users need to input age, batch, text size, number for, and upload a .csv file. The interface for listing models will display the models of the algorithms that were just executed. The prediction configuration interface allows users to input the model file for prediction, select the model for prediction, and upload it.

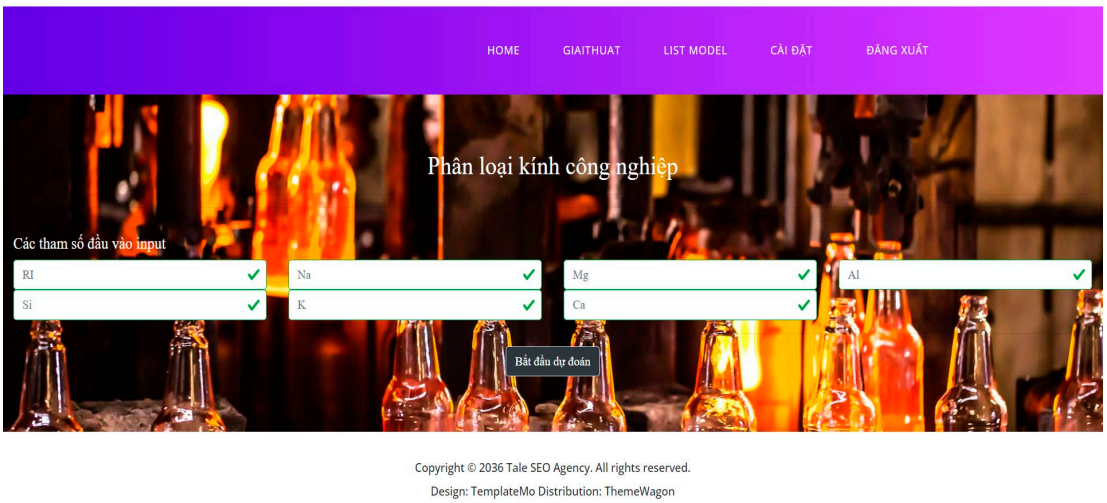

Figure 4. Predictive Configuration Page Interface

Allow users to input the initial parameters including RI, Na, Mg, Al, Si, K, Ca, and a start button for prediction initiation.

5. CONCLUSION

The topic has contributed to classifying different types of glasses based on users' required parameters. This helps minimize the amount of glass waste, bringing economic value to the people of Vietnam through the recycling and utilization of discarded glass, thereby reducing the significant amount of glass waste released into the environment. Enhancing the overall quality of life for the people of Vietnam and the world. This is a much-needed topic in Vietnam because glass recycling and classification are not yet widely practiced, so this topic contributes to approaching the task of classifying the recycling of glass by citizens, helping raise environmental awareness for each citizen, thereby bringing economic benefits to themselves, society, and the nation.

REFERENCES

[1] Rinkesh (2023). *Glass Recycling: Process of Recycling Glass and it's Benefits*. https://www.conserve-energy-future.com/recyclingglass.php

[2] Tạp chí môi trường. (2023). *Vì sao nên phân loại và tái chế kính thủy tinh*. http://tapchimoitruong.vn/giai-phap-cong-nghe-xanh-22/

[3] Okafor (2023). *Can you Recycle Glass? All about glass recycling*. https://www.trvst.world/waste-recycling/can-you-recycle-glass/

[4] Cleanipedia. (2023). *27 cách tái chế chai thủy tinh cũ thành đồ trang trí hữu ích tại nhà.* https://www.cleanipedia.com/vn/su-ben-vung/11-cach-tai-che-chai-thuy-tinh-thanh-vat-dung-lung-linh-xinh-xan.html

[5] Thu Ngân (2022). *Những ý tưởng tái chế chai thủy tinh sáng tạo, độc lạ*. https://www.btaskee.com/kinh-nghiem-hay/tai-che-chai-thuy-tinh/

[6] Experts. (2022). *Five tips for recycling glass*. https://www.hazardouswasteexperts.com/five-tips-for-recycling-glass/

[7] Momemtum recycling. (2023). *How is glass recycled*. https://utah.momentumrecycling.com/glass-recycling-process/

[8] Khoahoc.tv. (2023). *Thủy tinh có tái chế được không? Tái chế thủy tinh như thế nào?* https://khoahoc.tv/thuy-tinh-co-tai-che-duoc-khong-tai-che-thuy-tinh-nhu-the-nao-98475.

[9] Geeks (2021). *Multi-Layer Perceptron Learning in Tensorflow*. https://www.geeksforgeeks.org/multi-layer-perceptron- learning-in-tensorflow

Journal of Science and Development Economics
Nam Can Tho University

Website: jsde.nctu.edu.vn

Diagnosing the quality of wine using an adapting decision tree classifier for streaming data

Vo Ngoc Truong Duy[1], Vo Van Phuc[1], Tran Duy Khang[1], Ngo Ho Anh Khoi[1]
[1]*Faculty of Information Technology, Nam Can Tho University*

Received: November 10, 2023
Revised: November 30, 2023
Accepted: December 5, 2023

Keywords: *AI application, decision tree algorithm, wine quality diagnosis*

Từ khóa: *chuẩn đoán, chất lượng rượu vang, thuật toán cây quyết định, ứng dụng AI*

ABSTRACT

The research is focused on exploring the applications of Artificial Intelligence algorithms in handling diagnostic wine quality data. The article discusses the successful implementation of the Decision Tree algorithm for this purpose. This drives the main research goal, which revolves around integrating the Decision Tree with flexible sliding window techniques that can continuously adapt and update over time. The primary objective of the study is to address the wine quality diagnostic problem. Alongside this goal, there are additional smaller objectives to achieve. The initial step involves studying and researching theoretical foundations and measurement methods, as well as analyzing wine quality. Lastly, the goal of deploying a test application is set, aiming to create a Wine Quality Diagnostic Page. The interface of the page is designed to be user-friendly, intuitive, and informative about the functioning and content of the wine quality diagnostic method.

TÓM TẮT

Nghiên cứu tập trung vào việc khám phá các ứng dụng của thuật toán Trí tuệ nhân tạo trong việc xử lý dữ liệu chẩn đoán chất lượng rượu. Bài viết thảo luận về việc triển khai thành công thuật toán Cây quyết định cho mục đích này. Điều này thúc đẩy mục tiêu nghiên cứu chính xoay quanh việc tích hợp Cây quyết định với các kỹ thuật cửa sổ trượt linh hoạt có thể liên tục thích ứng và cập nhật theo thời gian. Mục tiêu chính của nghiên cứu là giải quyết vấn đề chẩn đoán chất lượng rượu vang. Bên cạnh mục tiêu này, còn có những mục tiêu nhỏ hơn cần đạt được. Bước đầu tiên bao gồm việc tìm hiểu, nghiên cứu cơ sở lý thuyết và phương pháp

đo lường cũng như phân tích chất lượng rượu. Cuối cùng, mục tiêu triển khai ứng dụng thử nghiệm được đặt ra nhằm tạo Trang Chẩn đoán Chất lượng Rượu. Giao diện của trang được thiết kế thân thiện với người dùng, trực quan và cung cấp nhiều thông tin về chức năng cũng như nội dung của phương pháp chẩn đoán chất lượng rượu.

1. INTRODUCTION

The issue at hand is understanding the significance of wine and why it is produced in various regions worldwide, including Vietnam. The state of wine in Vietnam: Despite the challenges posed by tropical viticulture, it is evident that quality wine can be produced in Vietnam. Vietnamese wine is crafted from the Cardinal grape variety, classified as a table grape in France. Additionally, there are a few Vitis Vinifera varieties - Cabernet Sauvignon, Chardonnay, Syrah - but they are relatively scarce. The importance of wine in Vietnam: Ladora Winery in Phat Chi – Da Lat, Lam Dong, has exclusively invested in a 6-hectare winery and a 20-hectare vineyard in Ninh Thuan to create a line of high-quality wines under the Vietnamese brand. The government and the Ministry of Tourism have also shown interest in investing more in this industry, as wine tourism is on the rise. Therefore, the significance of wine in Vietnam lies not only in its economic potential but also in its contributions to the domestic tourism sector.

According to various articles and international research reports, Sangodkar et al (2021) [1] explored the application of machine learning models for predicting wine quality; similarly, (Bhardwaj et al, 2022) [2], Piyush Bhardwaj introduces RF and AdaBoost models as machine learning classifiers to predict wine quality. The author evaluates these models based on accuracy, precision, recall, and is developing a web application based on machine learning for researchers and wine growers to predict wine quality using chemical and physical compounds present in their wines. Another study, by K. R. Dahal, author employs the Wine quality dataset from UCL to demonstrate the feasibility of using various statistical analyses to predict wine quality based on different parameters. This study implies that wine quality can be forecasted even before production, suggesting an alternative approach for understanding the variables influencing wine quality. In the Vietnamese context, (Hoàng Anh Lê, 2004) by Hoàng Anh Lê, the author focused on the significance of wine production. Moreover, in (Bùi Công Danh and Nguyễn Thị Diệu Hiền, 2021) [3] authors aims to analyze the importance of wine and the research timeline related to sensory evaluation, improvement, and new product creation. The dataset used in this research is named "Winequality_white.csv."All these works collectively share the primary objective of investigating wine quality prediction through the application of machine learning techniques. Decision Tree Analysis of Wine Quality Data was updated by Raj Parmar in 2019 (Parmar, 2019) [4]. The article focuses on utilizing the Decision Tree algorithm in a study related to wine quality diagnosis. The algorithm is implemented through a series of steps,

including library importation, dataset loading, data splitting into features and targets, division into training and testing sets, model construction, training on the training set, making predictions on the test set, evaluating model accuracy, and making predictions for new data. The model achieved a relatively high accuracy on the test set, indicating its effectiveness as a classifier. However, accuracy could be improved through adjusting model hyperparameters or employing different machine learning algorithms. Overall, the decision tree classifier is a powerful and versatile algorithm suitable for various classification tasks. The underlying datasets were downloaded from Kaggle. Assessing Wine Quality Using a Decision Tree was last updated by Seunghan Lee, and his colleagues in September 2015 (Lee et al., 2015) [5]. This article also focuses on the Decision Tree algorithm in the context of evaluating wine quality using decision trees. Wine quality assessment is crucial for the wine industry, and accurate evaluations are important for producers, distributors, and consumers. In recent years, decision trees have increasingly been used to predict wine quality ratings. This article summarizes Seunghan Lee's research on enhancing wine quality ratings using decision trees. The research is divided into three parts: research overview, research methodology, and significance and limitations of the study. The objective of Seunghan Lee's research is to improve wine quality ratings through the application of decision trees. Decision trees are a machine learning technique that employs a tree model to predict the value of a target variable based on multiple input variables. In this study, decision trees are utilized to predict wine quality ratings based on physical characteristics such as acidity, pH, alcohol concentration, and residual sugar.

The research is focused on exploring the applications of Artificial Intelligence algorithms in handling diagnostic wine quality data. The article discusses the successful implementation of the Decision Tree algorithm for this purpose. This drives the main research goal, which revolves around integrating the Decision Tree with flexible sliding window techniques that can continuously adapt and update over time. The primary objective of the study is to address the wine quality diagnostic problem. Alongside this goal, there are additional smaller objectives to achieve. The initial step involves studying and researching theoretical foundations and measurement methods, as well as analyzing wine quality. Lastly, the goal of deploying a test application is set, aiming to create a Wine Quality Diagnostic Page. The interface of the page is designed to be user-friendly, intuitive, and informative about the functioning and content of the wine quality diagnostic method.

2. MATERIALS AND METHODS

During the data exploration phase for the research topic, a multitude of datasets were discovered. However, three datasets exhibited the most comprehensive attributes and highest availability: Wine_Quality_Data by Ghassen Khaled (Khaled, 2023) [6], last updated on April 14, 2023, comprises 13 columns and 6497 rows of data. The data fields include various parameters like fixed acidity, volatile acidity, citric acid, residual sugar, chloride, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, alcohol content, quality, and label. This dataset is employed for training models

that classify white or red wines. Red Wine Quality from UCI Machine Learning, last updated in 2018, consists of 12 columns and 1599 rows of data. The data attributes encompass characteristics such as fixed acidity, volatile acidity, citric acid, residual sugar, chloride, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, alcohol content, and quality. This dataset is utilized to discern whether a wine is red or not. Wine Quality Data Set (Red & White Wine) by Ruthgn (2021) [7], updated in 2021, comprises 13 columns and 6497 rows of data. The attributes of this dataset are akin to the Wine_Quality_Data by Ghassen Khaled, differing primarily in parameter values and label positions. All of these datasets were found on Kaggle.

The enumerated datasets are all relevant to wine prediction, but only one dataset meets the requirements of the research topic. The other datasets present multiple issues that render them unsuitable for use in this study. For a dataset to be applicable, it must consist of numerical data, have specific class labels, and feature numerous attributes to yield objective outcomes. Among the listed datasets, only the Wine_Quality_Data fulfills these criteria. It provides the necessary capabilities for the research topic, making it the chosen dataset for this internship project.

This dataset may be temporary, as various new wine types have emerged, leading to potential changes in dataset classification in the future due to increased diversity. Hence, it can be identified as a dataset in an unstable environment. Consequently, the use of data classification for conventional machine learning algorithms within a static environment is not feasible. Old methods do not support classification and require advanced algorithms capable of processing data in dynamic environments. Currently, existing databases face the challenge of static concentration over time due to being trained using classical algorithms (in practice). This phenomenon occurs only once; when new data arrives, the previously learned information must be retrained entirely. For example, if data set 1 is used to create a model and new data set 2 arrives, data set 1 must be retrained from scratch alongside data set 2 to build a new model). Moreover, in the context of modern reality, where data environments evolve over time, training must occur continuously in real-time, and model predictions must be regularly updated. Consequently, data learning must transpire within an evolving data environment, which means that testing methods will continuously learn in a non-static environment. Several methods have been applied to transform classical algorithms into continuous learning approaches, replacing them with sliding window methods to evolve traditional machine learning techniques into advanced ones. The description of the Sliding Window-based approach is as follows:

Taking into account the evolution of concepts in an evolving data environment, the most recent training data is determined within a defined time window (either based on a time interval or several data instances). This approach can involve reclassifying "groups" (within the data selected by the temporary window) or updating the model if the online learning method permits. In this case, the process of "forgetting" (as mentioned above) is automatically managed by this learning method. This type of approach typically involves three steps: Detecting concept changes using

statistical tests across different windows. If an observed change occurs, select representative and recent data to adjust the models. Updating the models. The window size is predetermined by the user. The key point of these methods lies in determining the window size. Most methods employ a fixed-size window configured for each real-world problem. This way, classical algorithms can be applied in dynamic environments, but they lack the characteristics of progressive machine learning (they don't reuse stored data, only the model is used for improvement). Therefore, the historical part of the following algorithms focuses on presenting incremental machine learning algorithms, which have been researched and developed in recent years.

After transforming the data from its raw form to standardized format for use in the software, the dataset must be stored in a file with the ".csv" extension. Following all data normalization processes, we will obtain a standardized dataset consisting of 13 features, totaling 6497 data points, which encompass various parameters. Raw data must undergo preprocessing before being applied to the training process of the software. From these raw data, labels are transformed where the label "red," representing the color of red wine, is converted to the number "1," while the label "white," indicating the color of the second type of wine, is converted to "0." The remaining attributes used for assessment, such as fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulfates, alcohol, and quality, remain unchanged as they are already in numerical form and do not require conversion. These indices have been sourced from the aforementioned data and have been verified by domain experts. Each feature will have the following indices:

Fixed acidity: is a measure of non-volatile or non-volatile acids. These acids are derived from grapes. The main fixed acids found in wine are tartaric, malic, citric, and succinic acids. The unit of measurement is (g/L).

Volatile acidity: is a measure of the easily vaporizable (or gaseous) acids in wine. The primary volatile acid in wine is acetic acid, which is also the main acid associated with the aroma and taste of vinegar. The unit of measurement is (g/L).

Citric acid: is a weak organic acid, added to wine as a natural preservative and to enhance the acidity of the wine.

Residual sugar: is the amount of natural sugars left over from the grape after the fermentation process in red wine concludes. It is measured in (g/L).

Chlorides: In wine, the presence of 2 to 4 (g/L) of mineral acids, different from other organic acids, contributes to the potential salty taste of the wine. Thus, chlorides play a crucial role in the wine's saltiness. The unit of measurement is (g/L).

Free sulfur dioxide: is a measure of the amount of SO_2 that is not bound to other molecules and is used to calculate molecular SO_2. It is used during winemaking to prevent oxidation and the growth of microorganisms. The unit of measurement is (mg/L).

Total sulfur dioxide: is a measure of the combined and free forms of SO_2. Bound SO_2 refers to SO_2 molecules bound to other compounds, mainly aldehydes, pyruvates, and anthocyanins. It is used in winemaking to prevent oxidation and the growth of

microorganisms. Excessive levels of SO_2 can inhibit the fermentation process and result in undesirable sensory effects. The unit of measurement is (mg/L).

Density: Unit of measurement (g/cm^3).

pH: Describes the acidity or alkalinity of the wine on a scale from 0 (very acidic) to 14 (very alkaline); most wines fall within the pH range of 3-4.

Sulphates: Chemical formula $(SO_4)^{2-}$. It is an additive in wine that can contribute to the production of sulfur dioxide (SO_2) gas, acting as an antimicrobial and antioxidant agent. The unit of measurement is (g/L).

Alcohol: Percentage of alcohol content in the wine.

Quality: Rating scale from 1 to 10 based on sensory data, ranging from 3 to 8 in this dataset.

Predicted Label: This parameter carries a decision value and is particularly important. This label has only two possible values, either "0" or "1". If the wine is red, the label will be 1, while if the wine is white, the label will be 0. The prediction includes a total of 1599 data points for the red type and 4898 data points labeled as white.

The dataset used in this experiment consists of two parts: the training data portion includes 4545 data points (constituting 70% of the original data), and the testing data portion contains 1952 data points (representing 30% of the original data). The positions of these data points will vary in each experiment, and for each experiment, they will be randomly shuffled, both before training and after training.

The experiment will be conducted using an asynchronous batch learning model (Batch Learning) with a batch size of 4545. This means that the system will execute 4,545 steps, with approximately 129 data points per step. The model utilized in the experimental phase is a batch-wise data grouping model. This model will use the same dataset, organized into batches, by dividing the original dataset into smaller groups using 35 steps. Each batch of data contains 129 data points. This quantity strikes a balance between being not too large and not too small, facilitating the experimentation process.

All achieved results are based on Balanced Accuracy. Balanced Accuracy is a metric that can be used to evaluate the performance of a binary classifier. It is particularly useful when classes are imbalanced, meaning one of the two classes appears much more frequently than the other. Using Balanced Accuracy is significantly more complex than using regular Accuracy. Regular Accuracy simply computes the percentage ratio of a dataset's subgroup based on the total available data. Initially, this works well, but when the data is highly skewed (e.g. one data point in class A, 999 data points in class B), the accuracy calculation loses its correctness. To address this issue, a calculation method was devised, relying on true negative and true positive values, allowing for the computation of true negatives, true positives, false negatives, and false positives percentages. With all these parameters at hand, the formula for Balanced Accuracy can be employed to calculate the most accurate and optimal percentage truthfully. The Balanced Accuracy formula used in these experiments is as follows.

First, you need to refer to the confusion matrix:

	True Ground Truth Labels (1)	Incorrect Ground Truth Labels (0)
True positive predictions (1)	TP	FP
False positive predictions (0)	FN	TN

Figure 1. Confusion Matrix

In the matrix above, "positive" or "negative" in TP/FP/TN/FN refer to the predictions made, not the actual labels. (Thus, "false positive" is the case of incorrectly predicting positive). Below are the formulas for sensitivity and specificity based on the confusion matrix:

Sensitivity Formula: Sensitivity = TP/(TP + FN)

Specificity Formula: Specificity = TN/(TN + FP)

Balanced Accuracy Formula: $\text{Balanced accuracy} = \frac{(\text{Sensitivity} + \text{Specificity})}{2}$

The dataset used in the experiment consists of two parts: one for training and the other for testing. The training dataset comprises 1400 rows, while the testing dataset contains 701 rows. This dataset will vary in each experiment, but the data itself remains unchanged – only shuffling is applied to maintain data integrity. The model employed in the experiment follows a batch data approach. This model uses the same dataset in batches, achieved by partitioning the original dataset into smaller groups over 1400 steps. Each batch contains one data point, as this batch size is considered manageable and efficient in achieving optimal results for the task.

3. RESULTS AND DISCUSSION

By applying artificial intelligence techniques, specifically algorithms like Decision Trees, in combination with the Sliding Window method, a more equitable approach to assessing the authenticity of data can be achieved. Utilizing charts to compare the average outcomes of the three algorithms, this method ensures a fairer representation of data accuracy during the comparison of results from various algorithms. The experimental average results of the algorithms are depicted in the chart below (Figure 2).

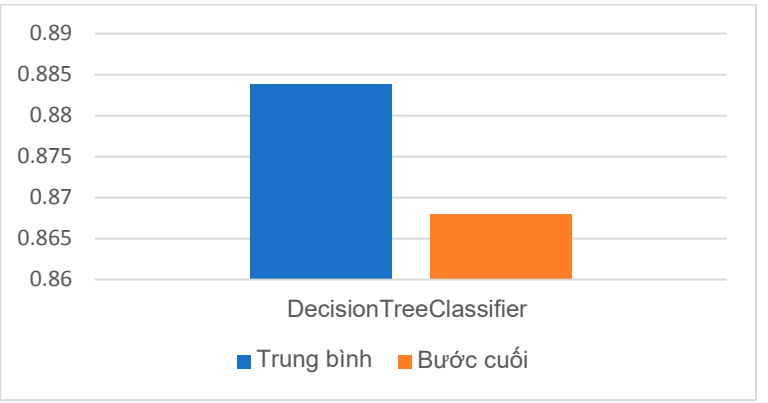

Figure 2. Chart of Average Percentage of Experimental Algorithm by Age (Decision Tree)

Looking at the data on the chart, we can analyze the average ratios and the final step's quite favorable ratio of the Decision Tree algorithm as follows:

The performance of the Decision Tree algorithm is remarkably stable, consistently achieving above 85% accuracy. The data illustrates that the Decision Tree algorithm maintains a consistent and noteworthy performance, with an average accuracy of over 85% (specifically 88.38%), and accuracy in the final stage also exceeding 85% (with an accuracy of 86.39%). This is a significant advantage, demonstrating the Decision Tree's capability to provide accurate predictions in the majority of cases, making it applicable to the wine quality diagnosis problem.

In addition to calculating the algorithm's average results, an alternative approach such as analyzing the experimental model results based on age reveals a more comprehensive, detailed perspective. This allows for a visual assessment that aids in arriving at the most accurate conclusions. The experimental model results based on age are depicted in the chart below (Figure 3).

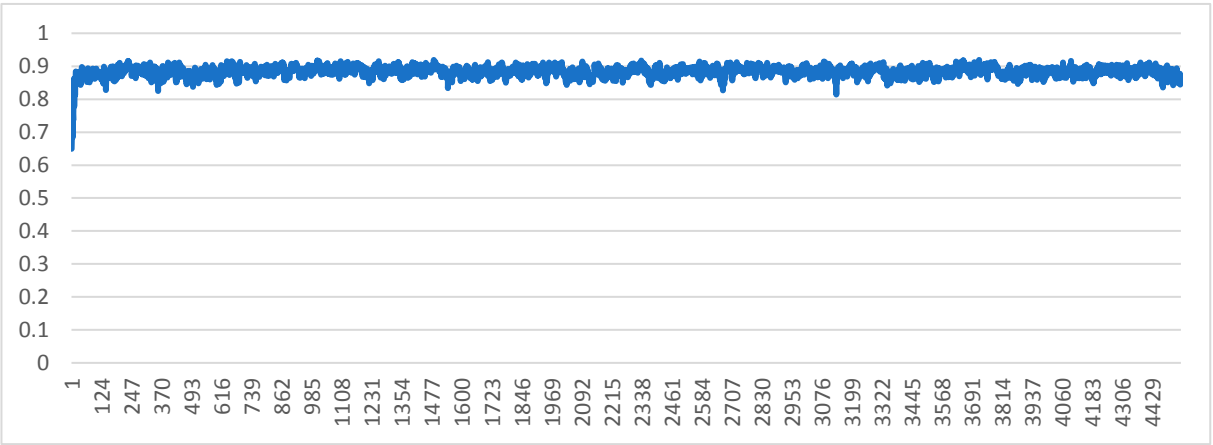

Figure 3. Chart of Experimental Model Results by Age (Decision Tree)

Examining the chart above, we observe that the Decision Tree algorithm starts with a relatively low point of 64% and gradually increases during the first 10 to 30 steps, after which it stabilizes. Looking at the graph, it's evident that the Decision Tree algorithm demonstrates reasonable stability. However, the highest accuracy rate of the Decision Tree can reach up to 91.72%, which is a notable figure when compared against various other algorithms. Specifically, during the phase from step 998 to 1008, as shown in the chart below (Figure 4).

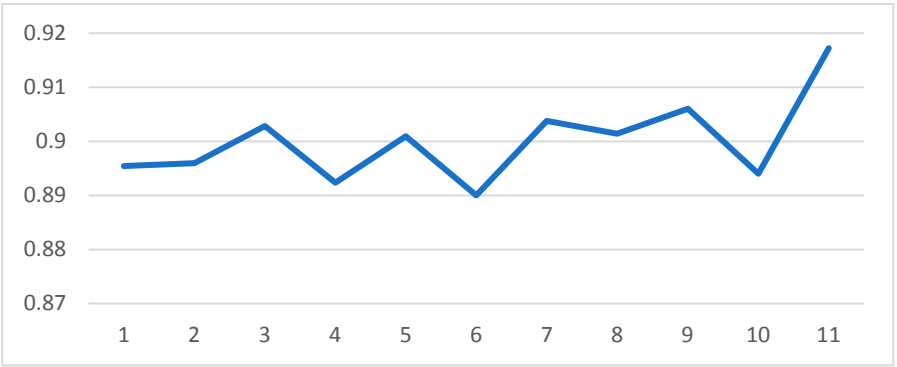

Figure 4. Peak Phase of the Decision Tree Algorithm (Batch 998-1008)

Upon examining the chart, it is evident that the Decision Tree consistently achieves a gradual increase, always surpassing the 80% mark. This is a rare occurrence when compared to other algorithms, as the Decision Tree demonstrates its ability to handle non-continuous and missing value data effectively. This capability minimizes the need for extensive data preprocessing and allows the algorithm to perform efficiently across various types of data. Despite encountering phases where results may not be optimal, the Decision Tree consistently stabilizes quickly, as depicted in the chart below (Figure 5).

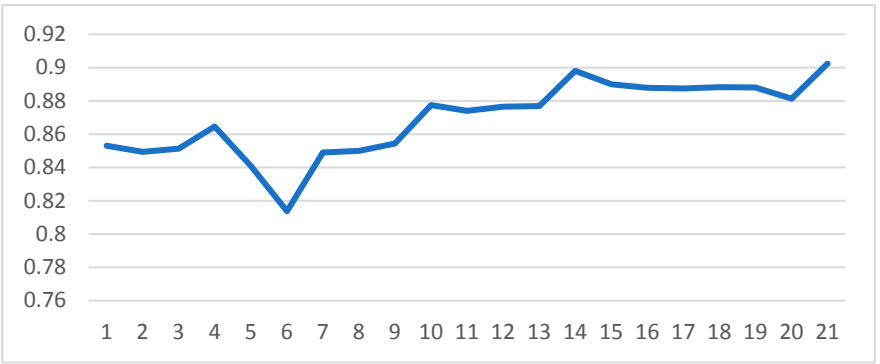

Figure 5. Lowest Dropped Phase of the Decision Tree (Batch 800 - 820)

At step 3135, the lowest accuracy rate achieved by the Decision Tree was 81.37%. However, by step 3139, the algorithm quickly rebounded to its stable performance with an accuracy of 87.74% and continued to gradually increase in subsequent steps. The accuracy rates during this phase do not vary significantly, ranging from 81.37% to 87.74%. The variation in performance is relatively small, indicating a consistent level of stability in the model.

In conclusion, the Decision Tree algorithm demonstrates numerous advantages and consistent performance with accuracy rates exceeding 85%. This underscores that the Decision Tree can be a solid and useful choice for various prediction and classification tasks. However, it's essential to note that each algorithm has its own strengths and weaknesses. The selection of an appropriate algorithm also depends on the specific requirements of the problem and the characteristics of the data.

Implementation of a real world application:

Based on the final results highlighted in the previous section, the chosen algorithm for addressing the problem is the Decision Tree algorithm. The project will encompass various functional nodes, including prediction functionality, execution of classical algorithms, a list of processed models, system configuration, and user authentication. This application will be implemented within a website environment, organized into two main user roles: Algorithm Setup (admin or developer) and Diagnosis User (end user). These roles are depicted in the use case diagram below: [Use Case Diagram illustrating the roles and functionalities of the application]**.** In this system, the Algorithm Setup role involves functions related to configuring and managing algorithms, while the Diagnosis User role focuses on utilizing the prediction capabilities and accessing processed models. The implementation will enable efficient interaction and utility for both types of users within the website framework.

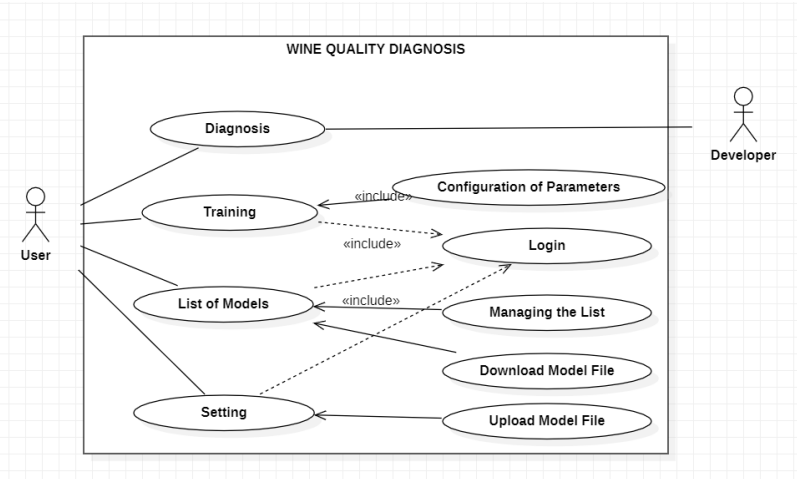

Figure 6. Use-Case Diagram of the System

To install the program, you need to first download the installation file named "StressDiagnostic.rar". After extracting the file, you will find a folder named "StressDiagnostic". To run the software, the user's computer needs to have certain Python libraries and Python 3.9.9 installed. Upon extraction, you will see a folder named "SETUP". Inside this folder, you will find a file named "python-3.9.9-amd64.exe", used to install Python 3.9.9, and a file named "inLib.bat" to install the necessary libraries required to run the software. Once the environment setup process is complete, a file named "Remove.bat" will be available. This file is used to delete unnecessary data files, including those used for testing purposes. It should be used in two scenarios: right after extraction and installation, and to remove all data from previous runs. To run the program, use the "Runserver.bat" file. This file is pre-configured to execute the command "py manage.py runserver", and the program will run on the default port "http://127.0.0.1:8000/". It's important to note that the user's computer should be connected to the internet at all times,

and the minimum system requirements include Windows 10, 2GB of RAM, and a 10GB or larger hard drive to ensure smooth and stable performance. After a successful installation, to use the software, access the port "http://127.0.0.1:8000" to reach the main page of the system. On the main interface page, a set of input fields will appear for data entry to perform predictions. Below are examples of the forms built within the "Wine Quality Diagnosis" system.

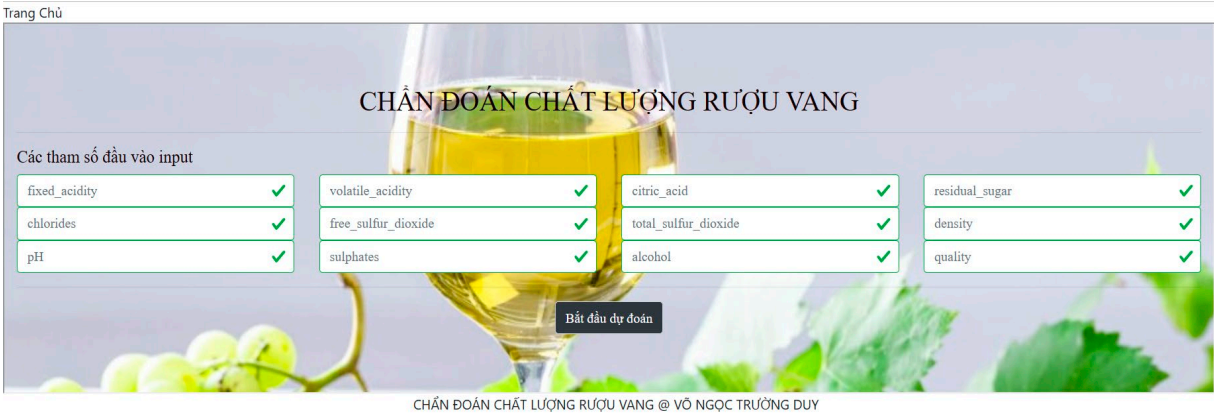

Figure 7. Main Interface of the Wine Quality Diagnosis System

5. CONCLUSION

Upon completing the research and report composition process, a comprehensive evaluation of the results can be presented. The report's content is clear and elaborates on data, charts, and algorithms specifically. The theoretical contribution is significant, as this project developed an algorithm with various training options. Simultaneously, the system was developed as a website interface, a rare feat achieved by very few systems. It tackled the challenge of dynamic data fluctuations by evolving the algorithm towards model-based training, whereas other algorithms usually only cater to static data through data-based training. However, this project only reached the research stage, leaving room for many future expansions. These may include implementing automated raw data processing within the system, optimizing model training processes, enhancing the system's interface for smoother user experience, refining the code for better aesthetics and broader user accessibility, and bringing the application into practical use, facilitating quick wine quality diagnoses for users. The research and development of this project were conducted in the context of the Vietnamese market, where there's a prevalence of counterfeit alcohol impacting both quality and human health. The system was built upon three Decision Tree algorithms as they fulfilled the requirements for changing data, as mentioned above.

REFERENCES

[1] Sangodkar, V.P. (2021). *Wine Quality Prediction Using Machine Learning.* https://www.ijraset.com/fileserve.php?FID=37629.

[2] Bhardwaj, P. (2022). *A machine learning application in wine quality prediction.* https://www.sciencedirect.com/science/article/pii/S266682702200007X.

[3] Bùi Công Danh, Nguyễn Thị Diệu Hiền (2021). *Đánh giá cảm quan rượu vang trắng bằng nổn nhân tạo.* https://sti.vista.gov.vn/tw/Lists/TaiLieuKHCN/Attachments/316303/CVv146S42021272.pdf.

[4] Parmar, R. (2019). *Decision Tree Analysis of Wine Quality Data,* https://www.kaggle.com/code/rajyellow46/decision-tree-analysis-of-wine-quality-data/notebook.

[5] Lee, S., Kang, K., & Park, J. (2015). *Assessing wine quality using a decision tree.* https://www.researchgate.net/publication/308862829_Assessing_wine_quality_using_a_decision_tree.

[6] Khaled, G. (2023). *Wine Quality Data*. https://www.kaggle.com/datasets/ghassenkhaled/wine-quality-data/discussion.

[7] Ruthgn (2021). *Wine Quality Data Set (Red & White Wine).* https://www.kaggle.com/datasets/ruthgn/wine-quality-data-set-red-white-wine?resource=download.

Journal of Science and Development Economics
Nam Can Tho University

Website: jsde.nctu.edu.vn

Heart disease prediction using multilayer perceptron in a dynamic environment

Le Thi My Nhu[1], Ngo Ho Anh Khoi[1], Duong Duy Khanh[2]
[1]*Faculty of Information Technology, Nam Can Tho University*
[2]*Hanoi Tam Anh Hospital*

Received: November 10, 2023
Revised: November 30, 2023
Accepted: December 5, 2023

Keywords: *cardiovascular diseases, heart failure prediction, multilayer perceptron*

Từ khóa: *bệnh tim mạch, cảm biến đa lớp, dự đoán suy tim*

ABSTRACT

In recent years, the incidence and mortality rates due to cardiovascular diseases have been on the rise globally. This is the primary reason why the main objective of this topic is to investigate techniques aimed at solving the problem of heart disease diagnosis. The research methodology for this topic involves the use of the scientific experimental approach, conducted on the Multilayer Perceptron (MLP) algorithm using the Heart Failure Prediction Dataset as the foundational dataset. This research addresses a highly significant societal issue. If further studied and developed, it has the potential to empower individuals to proactively and effectively prevent heart diseases. The prediction of heart disease has become a crucial field of study, aiding in early detection, risk assessment, and the implementation of preventive measures. This article summarizes several important aspects related to heart disease prediction based on scientific machine learning methods.

TÓM TẮT

Trong những năm gần đây, tỷ lệ mắc và tử vong do bệnh tim mạch ngày càng gia tăng trên toàn cầu. Đây là lý do chính tại sao mục tiêu chính của chủ đề này là nghiên cứu các kỹ thuật nhằm giải quyết vấn đề chẩn đoán bệnh tim. Phương pháp nghiên cứu cho chủ đề này bao gồm việc sử dụng phương pháp thực nghiệm khoa học, được thực hiện trên thuật toán Perceptron đa lớp (MLP) sử dụng bộ dữ liệu dự đoán suy tim làm bộ dữ liệu cơ bản. Nghiên cứu này đề cập đến một vấn đề xã hội rất có ý nghĩa. Nếu được nghiên cứu và phát triển thêm, nó có khả năng trao quyền cho các cá nhân ngăn ngừa bệnh tim một cách chủ động và hiệu quả.

Dự đoán bệnh tim đã trở thành một lĩnh vực nghiên cứu quan trọng, hỗ trợ phát hiện sớm, đánh giá rủi ro và thực hiện các biện pháp phòng ngừa. Bài viết này tóm tắt một số khía cạnh quan trọng liên quan đến dự đoán bệnh tim dựa trên phương pháp học máy khoa học.

1. INTRODUCTION

In recent years, cardiovascular diseases have been confirmed as the leading cause of death globally, including in Vietnam. Heart diseases often progress silently, with many cases going undetected as they do not cause significant pain. Individuals with heart conditions might continue working and engaging in normal activities, leading to complacency. Older adults are particularly vulnerable to heart diseases. The elderly population is susceptible due to gradual damage to blood vessels over time, often resulting in the accumulation of arterial plaques. Recommendations include implementing comprehensive strategies and approaches across the entire population to address the burden of heart disease in Vietnam. Health education, increasing awareness, and modifying behaviors among heart disease patients are crucial for early detection, prevention, and timely treatment. These efforts can help slow down the progression of heart diseases, improve the quality of life for affected individuals, and simultaneously alleviate economic burdens on families and society as a whole.

In the article by Schocken and colleagues, the increasing prevalence of heart failure worldwide, including in developing regions like Vietnam, poses significant challenges for caregivers, researchers, and policymakers. Consequently, prioritizing the prevention of this global catastrophe is of utmost importance. In the article by Ashrafian and colleagues, the authors discuss how neural resistance has successfully reduced the incidence and mortality rates of heart failure. Further discussions delve into therapies based on new mechanisms to enhance metabolic processes and insulin resistance in heart failure (Ashrafian et al., 2007) [1]. In the report by Virani and collaborators, the focus is on the American Heart Association's collaboration with the National Institutes of Health to provide an annual updated statistical report on heart disease, stroke, and cardiovascular risk factors. This statistical update presents the latest data on a range of cardiovascular conditions including stroke, congenital heart disease, arrhythmias, atherosclerosis, coronary artery disease, heart failure, and valve diseases. (Virani et al., 2020) [2]. Trang (2018) [3] focused on assessing clinical and preclinical characteristics in patients admitted for acute heart failure due to localized coronary ischemia. The study investigates factors promoting acute events and their correlations with short-term outcomes. Lastly, Vu Thi Thom addresses the rising incidence and mortality rates of coronary heart disease in Vietnam. Thus, they conducted a cross-sectional descriptive study on 269 university staff members in Hanoi in 2016, aged 20 to 64. The study reveals higher risk factor rates in males, with the university staff having lower rates of hypertension and dyslipidemia compared to the general community. Gender

was found to be correlated with overweight and obesity, high blood pressure, and dyslipidemia. (Vu Thi Thom, 2018) [4].

In Colombet (2000) [5], the authors conducted a comparison of three algorithms, CART, Multilayer Perceptron, and logistic regression, to predict cardiovascular risk from real-world data. Estimating multivariate risk is currently required in cardiovascular disease prevention guidelines. Limitations of existing statistical risk models have led to the exploration of machine learning methods. The research data was randomly divided into a training set (n=10,296) and a testing set (n=5,148). The accuracy results for the three algorithms were as follows: 65.9%, 76.0%, 69.1%. In Yan et al (2006) [6], the authors studied a Multilayer Perceptron-based decision support system developed to aid in heart disease diagnosis. The system's input layer comprises 40 input variables categorized into four groups, encoded using proposed encoding schemes. The number of nodes in the hidden layer is determined through layer-wise learning. Each of the 5 nodes in the output layer corresponds to a specific heart disease. In cases of missing patient data, the system employs mean replacement for handling missing values. Furthermore, an enhanced backpropagation algorithm is utilized for system training. A total of 352 medical records were gathered from patients with 5 heart diseases, used for system training and testing. Cross-validation, holdout, and bootstrapping methods were applied to evaluate the system's generalization. The results demonstrate that the proposed MLP-based decision support system achieves highly accurate diagnosis (>90%) with relatively low time overhead (<5%), showcasing its utility in aiding the heart disease diagnostic process.

2. MATERIALS AND METHODS

During the process of data exploration for the research topic, numerous datasets were found (approximately 1163 datasets related to heart disease issues). However, out of these, only four datasets exhibited the highest availability and comprehensive parameter details, including:

- Heart Failure Prediction Dataset by Fedesoriano: This dataset contains 918 rows and 12 columns, encompassing variables such as age, gender, chest pain type, resting blood pressure, maximum heart rate, cholesterol level, etc. (Fedesoriano, 2021) [7].
- Heart Failure Prediction by Larxel (2022) [8]: With 299 rows and 13 columns, this dataset includes variables like age, gender, smoking duration, ischemia status, etc.
- Heart Disease Dataset by Yasser (2021) [9]: Consisting of 304 rows and 14 columns, this dataset covers variables like age, gender, chest pain type, resting blood pressure, exercise-induced chest pain, etc.
- Heart Diseases by Kakaraparthi (2022) [10]: This dataset comprises 303 rows and 14 columns, featuring variables like age, gender, chest pain type, resting blood pressure, maximum heart rate, cholesterol level, major vessels count, etc. Although these datasets are all related to predicting heart diseases, only one dataset meets the requirements of the project due to issues in other datasets such as insufficient data or outdatedness. The dataset must be numerical, have specific class labels, and contain multiple fields to provide objective outcomes. The only dataset meeting these criteria is the Heart Failure Prediction Dataset.

Current databases face a significant issue in adapting to changing data over time. Classical algorithms, on which existing databases are often trained, can only be trained once and need to be re-learned from scratch when data changes. In modern data environments, data evolves continuously, necessitating quick adaptation to these changes. To address this, continuous learning in a non-stable environment has been proposed as an alternative solution. Continuous learning allows databases to learn data continuously in a changing environment, facilitating updates and adjustments to their predictive models. This enhances the database's adaptability to data changes and improves the accuracy of their predictive models.

The term "concept drift" has been widely used in problems in dynamically changing environments like heart prediction. In fact, the concept of drift forms the basis for gradual changes, continuous shifts, and the "forgetting" of previous situations. However, in dynamically changing environments, the complexities often result in scenarios that change rapidly, slowly, or even involve the reappearance of previously disappeared knowledge. In such intricate situations, the notion of "dual dilemma" regarding stability or adaptability gains its full significance. It's important to note that these approaches aren't universally suitable and vary based on real-world conditions.

The "Sliding Windows" approach is one of three methods proposed to address concept drift. This method involves considering the evolution of concepts in a non-stable environment by using a "sliding window" as seen in the FLORA approach. The principle involves updating the model at each time step using the most recent training data defined by a window of time or data count. This approach can reclassify data into "groups" (based on temporary window data) or update models if online learning methods permit. The challenge lies in determining the window size, which is often fixed for practical applications. The dataset is temporary, reflecting diverse emerging standards, potentially leading to changes in dataset classification and indicating an unstable environment. Consequently, traditional algorithms designed for static environments aren't viable. Incremental learning algorithms that accommodate dynamic environments must be used.

Considering the mentioned characteristics and drawing from historical research on various methods, the upcoming research will employ the Sliding Windows method, which is the most suitable approach. This method will be combined with the Multilayer Perceptron algorithm.

The Heart Failure Prediction Dataset is provided by the author FEDESORIANO, who is a member of the Kaggle website. The dataset was last updated on September 11, 2022, according to the Vietnam time zone. Upon downloading, the dataset is in raw form, which cannot be directly used for this study. It requires substantial transformation to convert various parameters and characteristics into a usable format. Raw data needs preprocessing before it can be applied to the software training process.

The following transformations are performed:

- Sex Attribute: M representing male is converted to "1", and F representing female is converted to "0".
- ExerciseAngina Attribute: Y is changed to "1", N is changed to "0".

- ChestPainType Attribute: ASY is converted to 1, TA to 2, ATA to 3, NAP to 4.
- FastingBS Attribute: If fasting blood sugar > 120, it's converted to "1", otherwise "0".
- RestingECG Attribute: Normal is changed to 1, ST to 2, LVH to 3.
- ST_Slope Attribute: Up is changed to 1, Flat to 2, Down to 3.

Other attributes (Age, RestingBP, MaxHR, Oldpeak, Cholesterol) are retained as they are numeric. After these transformations, the dataset must be saved in ".csv" format. Each row should be entered in a single cell, with parameters and labels separated by commas. Parameters should be placed before the label as per system regulations.

The processed dataset will have 12 features, totaling 918 data entries, with attributes like Age, Sex, ChestPainType, RestingBP, Cholesterol, FastingBS, RestingECG, MaxHR, ExerciseAngina, Oldpeak, ST_Slope, and Prediction Label.

For the experimental setup, the dataset is divided into two parts: a training dataset (642 data entries, 70% of the original data) and a testing dataset (276 data entries, 30% of the original data). The positions of these data entries are shuffled randomly for each experiment and are reshuffled after training.

The experiment is conducted through an indirect experimental model (Batch Learning, batch size = 641), where each step involves one data entry. Increasing the batch size equals 70% of the original data and helps improve accuracy, particularly for large, complex datasets. However, it's time-consuming.

All achieved results are based on Balanced Accuracy, which is particularly useful for evaluating binary classifier performance in imbalanced classes. It's more complex than traditional Accuracy calculations. Balanced Accuracy addresses the issue of accuracy skewing when classes are imbalanced (e.g. one data entry in class A, 999 in class B). It calculates percentages based on true negatives, true positives, false negatives, and false positives.

The formula for Balanced Accuracy used in these experiments is:

First, reference the confusion matrix:

	True Ground Truth Labels (1)	**Incorrect Ground Truth Labels (0)**
True positive predictions (1)	TP	FP
False positive predictions (0)	FN	TN

It's important to have:

Correct, TPR (True Positive Rate) is also known as the sensitivity or recall, and it represents the ratio of true positive predictions (correctly detecting the positive class) to the actual positive instances in the dataset. It's calculated using the formula: TPR = $\frac{TP}{TP+FN}$

The TNR (True Negative Rate) is also known as specificity and represents the ratio of true negative predictions (correctly detecting

the negative class) to the total actual instances of the negative class in the dataset. It's calculated using the formula: $\text{TNR} = \frac{TN}{TN+FP}$

After obtaining these two metrics, Balanced Accuracy is calculated using the formula: $Balance\ Accuracy = \frac{TPR+TNR}{2}$

3. RESULTS AND DISCUSSION

Using artificial intelligence methods, specifically algorithms like the Multilayer Perceptron, in combination with the Sliding Window technique, we have applied a comparative approach. Utilizing charts, we compare the average results of the algorithm. The experimental average results of the algorithm are represented in the chart below (Figure 1).

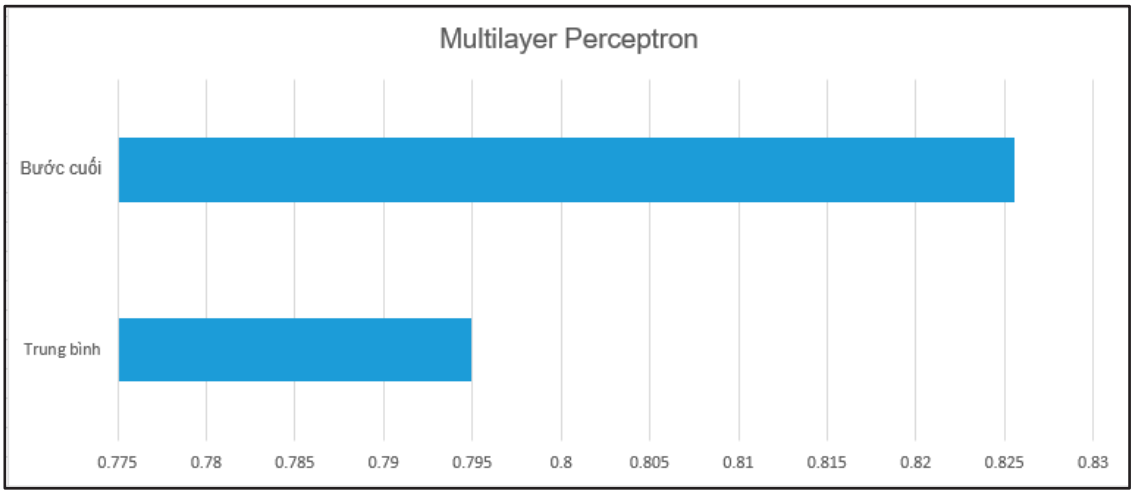

Figure 1. Chart comparing the average percentage and final step of the Multilayer Perceptron algorithm

Analyzing the data from the chart, we can deduce the following insights about the average percentage and the final step of the Multilayer Perceptron algorithm:

The data illustrates that the Multilayer Perceptron achieves a stable and remarkable performance, with an accuracy rate at the final step exceeding 80% (specifically 82.55%), and an average accuracy of 79.49%. Apart from comparing the accuracy percentage between the average and final steps, it's also possible to assess the algorithm in a more comprehensive and detailed manner by examining each step. This allows for a clearer evaluation of the algorithm's progress through the accuracy progression chart of the Multilayer Perceptron presented below (Figure 2).

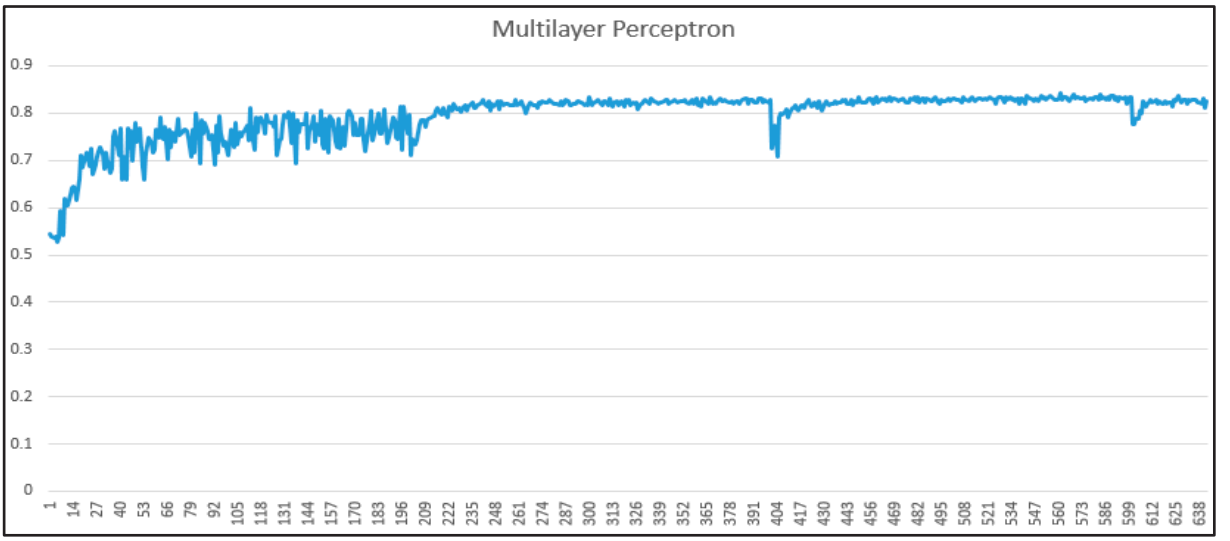

Figure 2. The accuracy progression chart of the Multilayer Perceptron algorithm

Observing the accuracy progression chart of the Multilayer Perceptron algorithm, we notice that the algorithm starts with a relatively low point, reaching only 55.45% accuracy. However, the stability has gradually increased over time. Specifically, at the beginning, the algorithm achieves an average accuracy of 55.45%, but by step 9, it has risen to 61.80%. Subsequently, there is a continuous increase in accuracy over the following steps, maintaining stability. Looking at the chart, we can see that the Multilayer Perceptron reaches its highest ratio at step 561, achieving an accuracy rate of 84.11%, which is quite remarkable. It's evident that the greatest growth in accuracy occurs in the phase from step 546 to 561, as illustrated in the graph below (Figure 3).

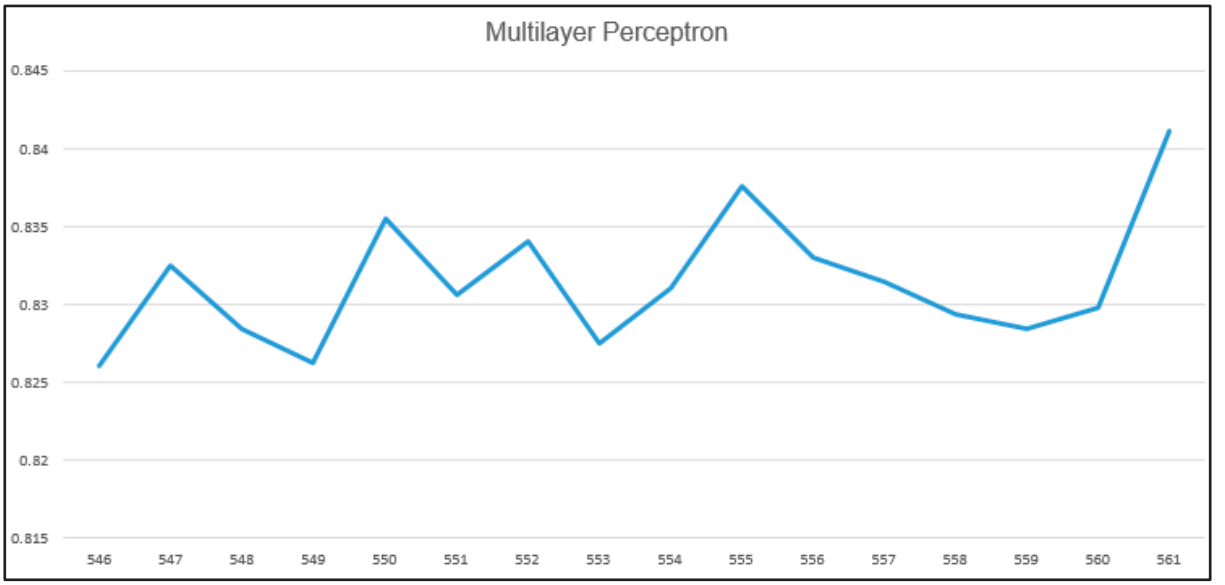

Figure 3. The highest level stage of the Multilayer Perceptron algorithm (Batch 546-561)

Examining the chart, we can clearly observe the steady and consistent increase of the Multilayer Perceptron algorithm's performance, consistently reaching above the 80% mark. The Multilayer Perceptron consistently proves itself to be a solid algorithm, even though there are some stages where the results are not favorable. However, the algorithm quickly manages to achieve a good state and maintains stability, as depicted in the graph below (Figure 4).

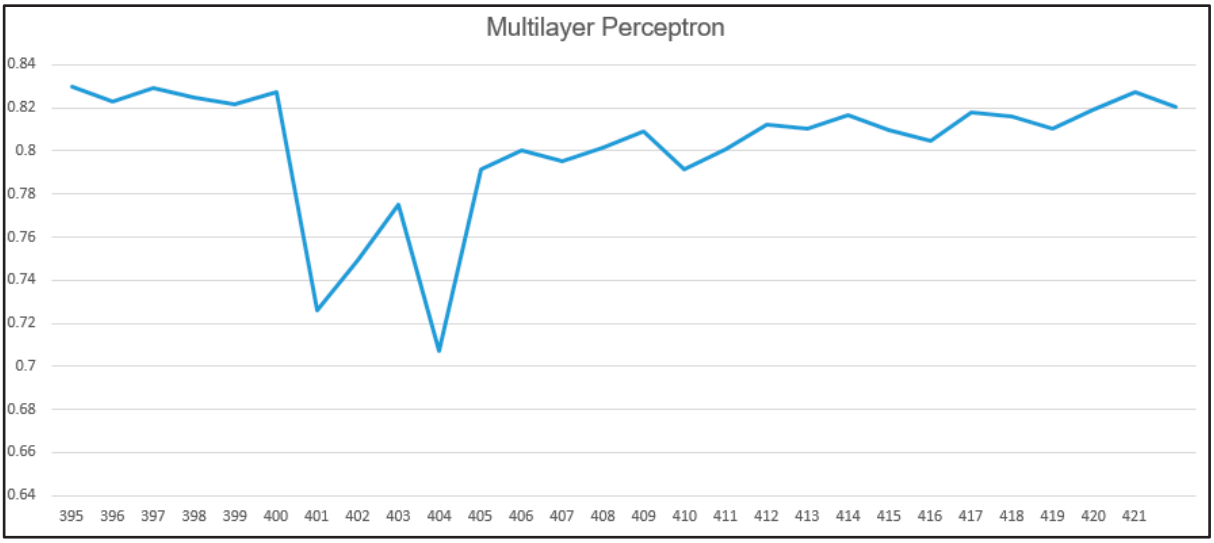

Figure 4. The lowest level stage of the Multilayer Perceptron algorithm (Batch 623-642)

In steps 401-405, the lowest percentage range of the Multilayer Perceptron algorithm fluctuated between 72.61% and 79.17%. However, at step 406, the algorithm quickly returned to its stable performance, achieving 80.03%. In the subsequent steps, although there were fluctuations in the percentage between the steps, the differences were not substantial. Overall, the Multilayer Perceptron algorithm showcases several advantages and relatively consistent performance with an accuracy rate above 80%. This indicates that the Multilayer Perceptron can be a valuable and useful choice in various prediction tasks.

Demo installation:

Each function of the system has been completed and fulfills the initial requirements. The system is divided into distinct sections for regular users and developers. Regular users have access to a form for predicting cardiac diseases, while developers have additional forms to work within the system, including a login form, model list form, data training form, and settings form. The system holds a high potential for practical application due to the increasing prevalence of cardiac issues in Vietnam, driven by unhealthy lifestyles and habits. Moreover, diagnostic standards for cardiac diseases are subject to change over time due to various objective reasons. Consequently, the dataset will undergo significant changes, and the Multilayer Perceptron algorithm is fully equipped to adapt to these modifications. There are two primary functions: Algorithm Setup (admin or developer) and Diagnostics (user).

Diagnosis User Interface:

The homepage of this system is a heart disease prediction page, constructed with the following input parameters: Age, Sex, ChestPainType, RestingBP, Cholesterol Level, Fasting Blood Sugar Level, Resting Electrocardiographic Results, Max Heart Rate Achieved, Exercise-Induced Angina, ST Depression Induced by Exercise Relative to Rest, and ST Slope. Each parameter is of the floating-point data type. On the diagnosis form designed for users, there will be a button labeled "Start Prediction" and 11 input fields for entering diagnosis data.

Figure 5. Interface for Prediction Users

In this project, the system has been packaged as a ZIP file named "ChanDoanBenhTim-main.zip". After users download and extract the file, there will be a directory named "ChuanDoan". Inside this directory, there are files to run the program. The system requires your computer to be constantly connected to the internet and has the following minimum configuration: Windows 10, 2GB RAM, and 10GB of hard drive space. To run the software, follow these steps:

Python Installation: Navigate to the "SETUP" directory and run the file "python-3.9.9-amd64.exe" to install Python version 3.9.9. Library Installation: Run the file "CaiThuVien.bat" in the same directory to install the required libraries. Start the Server: Run the file "RunServer.bat" to start the program. Access the Application: Open your web browser and access the address 'http://127.0.0.1:8000/'. The program will be running on the default port '8000'. By following these steps, you will be able to install and run the heart disease diagnosis software.

5. CONCLUSION

The completion of the research project as well as the report writing marks a comprehensive achievement. Concerning the software aspect, the Multilayer Perceptron algorithm has been successfully integrated into the program, enhancing diagnostic accuracy and simultaneously addressing dynamic data

processing challenges that other algorithms have yet to tackle. A user interface has been developed, implementing the software through two approaches: command line and graphical interface. In terms of the report, all initially outlined chapters have been satisfactorily completed, with clear content elucidating data, charts, and algorithms in a specific manner. This project has only advanced to the research phase; therefore, there remain numerous opportunities for expansion in the future. These include incorporating automated raw data processing directly within the system, optimizing model training processes, refining the user interface for smoother operation, transitioning the application to practical use, facilitating swift cardiac disease diagnoses, and further diversifying functionality - potentially extending it to mobile platforms.

REFERENCES

[1] Ashrafian, H., Michael, P., Frenneaux, M.D., & Opie, L.H. (2007). *Metabolic Mechanisms in Heart Failure.* https://www.ahajournals.org/doi/epub/10.1161/CIRCULATIONAHA.107.702795

[2] Virani, S. S., Alonso, A., Benjamin, E. J., Bittencourt, M. S., Callaway, C. W., & Carson, A. P. (2020). Heart disease and stroke statistics 2020 update: a report from the American Heart Association. *Circulation, 141*(9), e139-e596.

[3] Trang, T. T. H., Phong, P. Đ., & Hạnh, V. Đ. (2018). Nghiên cứu một số yếu tố thúc đẩy suy tim cấp và biến cố ngắn hạn ở bệnh nhân suy tim mạn tính do bệnh tim thiếu máu cục bộ. *Tạp chí Tim mạch học Việt Nam*, (84+ 85), 138-144.

[4] Vu Thi Thom, Vu Van Nga, Do Thi Quynh, & Vu Thi Mai Anh (2018). Các yếu tố nguy cơ mắc bệnh tim mạch của nhân viên một trường đại học tại Hà Nội. *Tạp chí Khoa học ĐHQGHN: Khoa học Y Dược*, Tập 34, Số 2 (2018) 89-96

[5] Colombet, I., Ruelland, A., Chatellier, G., Gueyffier, F., Degoulet, P., & Jaulent, M. C. (2000). Models to predict cardiovascular risk: comparison of CART, multilayer perceptron and logistic regression. In *Proceedings of the AMIA Symposium* (p. 156). American Medical Informatics Association.

[6] Yan, H., Jiang, Y., Zheng, J., Peng, C., & Li, Q. (2006). A multilayer perceptron-based medical decision support system for heart disease diagnosis. *Expert Systems with Applications*, *30*(2), 272-281.

[7] Fedesoriano (2021). *Heart Failure Prediction Dataset.* https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction.

[8] Larxel (2022). *Heart Failure Prediction*. https://www.kaggle.com/datasets/andrewmvd/heart-failure-clinical-data.

[9] Yasser, H (2021). *Heart Disease Dataset.* https://www.kaggle.com/datasets/yasserh/heart-disease-dataset.

[10] Kakaparthi, C. (2022). *Heart_Diseases*. https://www.kaggle.com/datasets/charankakaraparthi/heart-dissease.

Journal of Science and Development Economics
Nam Can Tho University

Website: jsde.nctu.edu.vn

Predicting edible and toxic mushrooms with multi-layer perceptron method in streaming data

Nguyen Ngoc Pham[1], Phan Thi Xuan Trang[1], Tran Thi Thuy[2], Ngo Ho Anh Khoi[1]
[1]*Faculty of Information Technology, Nam Can Tho University*
[2]*Nam Can Tho University*

Received: November 10, 2023
Revised: November 30, 2023
Accepted: December 5, 2023

Keywords: *advanced machine learning, artificial intelligence, mushroom diagnosis, MLP classifier*

Từ khoá: *chẩn đoán nấm, học máy nâng cao, phân loại MLP, trí tuệ nhân tạo*

ABSTRACT

In today's rapidly advancing technological landscape, a significant portion of current technological developments is centered around the field of Artificial Intelligence (AI). Machine Learning, a subfield of AI, applies statistical and mathematical methods to enhance computer performance. AI has made substantial contributions to solving a wide range of problems over the past decade. Particularly, given the current context where distinguishing between different types of mushrooms is not well understood, especially to prevent the consumption of poisonous mushrooms that can have severe health consequences. The chosen algorithm for this project is the MLP (Multi-layer Perceptron) classifier, which has seen significant development in recent years and has widespread applications in various AI domains. This is why it has been selected as the foundation for the system in this project. The research topic aims to utilize this algorithm to construct a system that rapidly predicts whether a mushroom is edible or poisonous, facilitating practical applications. The application of artificial intelligence in the development of a system related to cognitive differentiation holds the promise of reducing the use of toxic mushrooms in daily life.

TÓM TẮT

Trong bối cảnh công nghệ phát triển nhanh chóng ngày nay, một phần đáng kể của sự phát triển công nghệ hiện nay tập trung vào lĩnh vực Trí tuệ nhân tạo (AI). Machine Learning, một lĩnh vực con của AI, áp dụng các phương pháp thống kê và toán học để nâng cao hiệu suất máy tính. AI đã có những đóng góp đáng kể

trong việc giải quyết nhiều vấn đề trong thập kỷ qua. Đặc biệt, trong bối cảnh hiện nay, việc phân biệt giữa các loại nấm khác nhau chưa được hiểu rõ, đặc biệt là để ngăn chặn việc tiêu thụ nấm độc có thể gây hậu quả nghiêm trọng cho sức khỏe. Thuật toán được chọn cho dự án này là thuật toán phân loại MLP (Multi-layer Perceptron), thuật toán này đã có sự phát triển đáng kể trong những năm gần đây và có ứng dụng rộng rãi trong nhiều lĩnh vực AI khác nhau. Đây là lý do tại sao nó được chọn làm nền tảng cho hệ thống trong dự án này. Đề tài nghiên cứu nhằm mục đích sử dụng thuật toán này để xây dựng một hệ thống dự đoán nhanh chóng nấm ăn được hay nấm độc, tạo điều kiện thuận lợi cho các ứng dụng thực tế. Việc ứng dụng trí tuệ nhân tạo vào việc phát triển hệ thống liên quan đến sự khác biệt về nhận thức hứa hẹn sẽ giảm việc sử dụng nấm độc trong cuộc sống hàng ngày.

1. INTRODUCTION

In recent years, the cultivation of edible mushrooms has experienced robust growth in various regions across my country, becoming a relatively significant sector within agricultural production. Due to its high-profit potential, mushroom cultivation has emerged as a rich source of nutrition for humans, particularly in the case of certain medicinal mushroom varieties. Furthermore, the cultivation of straw mushrooms accelerates the process of agricultural and forestry waste decomposition, contributing to environmental cleanliness. Moreover, the practice of growing edible mushrooms enhances the value per unit of cultivated land area, generating employment opportunities and playing a role in adjusting the structure of the agricultural and rural economy. It also provides valuable export products, augmenting income for farmers and improving their livelihoods. Additionally, it aids in rural development by contributing to the overall economic transformation and providing sustainable growth. The cultivation of mushrooms not only offers economic benefits but also plays a pivotal role in waste management and ecological balance. It signifies a harmonious interplay between agriculture and the environment, showcasing how modern agricultural practices can positively impact both human well-being and the ecosystem. This shift towards mushroom cultivation exemplifies the dynamic evolution of agricultural practices and the recognition of mushrooms as a valuable resource, fostering socio-economic development while promoting a sustainable and ecologically conscious approach to farming. In the present global context, an exhaustive examination encompassing 2,786 mushroom species from 99 countries was conducted. This comprehensive review drew upon 9,783 case reports sourced from over 1,100 references. The analysis yielded valuable insights, identifying 2,189 mushroom species as edible, with a safe consumption status confirmed for 2,006 of them. Furthermore, 183 species were identified as requiring some form of pre-treatment before safe consumption, or they were associated with

allergic reactions in certain individuals. Additionally, there were 471 species for which edibility remained uncertain due to incomplete or missing consumption data, and 76 species whose edibility and toxicity status remained unconfirmed, with varying opinions on their safety. These findings underscore the critical importance of exercising caution when foraging for mushrooms and highlight the necessity for robust policies and guidelines to ensure the safe utilization of mushrooms on a global scale. It's worth noting that many regions across the world have implemented management and control measures to address mushroom utilization in light of these findings (Li, 2021) [1]. The situation in Vietnam is characteristic of a country with a rich diversity of mushroom species. However, approximately 200 mushroom species in Vietnam are considered toxic and potentially hazardous to human health. Some common poisonous mushroom types in Vietnam include Thanh Ba toxic mushroom, water caltrop toxic mushroom, and Amanita toxic mushroom. Toxic mushrooms in Vietnam are typically found during the rainy season, when the weather is humid. In many cases, Vietnamese individuals have experienced poisoning after consuming mushrooms of uncertain origin or improperly prepared ones (Nguyen Thi Thu Ha, 2016) [2]. In the research of Tutuncu updated in 2022 (Tutuncu, 2022) [3], based on 22 features within the Mushroom dataset and four different machine learning algorithms, the outcomes of applying these algorithms to the mushroom dataset are compared. Models were constructed to classify mushrooms into edible and poisonous categories. The success rates of these models were obtained from Naive Bayes, MLP classifier, Support Vector Machine, and AdaBoost algorithms, with success rates of 90.99%, 98.82%, 99.98%, and 100% respectively. Upon further examination of these results, considering external mushroom characteristics, it was determined whether a mushroom is edible or poisonous with 100% accuracy using the AdaBoost model. This study highlights the significant capabilities of machine learning algorithms in accurately classifying mushrooms based on their features, contributing to the vital task of distinguishing between edible and toxic varieties for human safety.

Based on the aforementioned article, a study was conducted to explore the application of modern Artificial Intelligence algorithms in solving a mushroom diagnosis dataset, with a specific focus on the MLP classifier (Multi-layer Perceptron). Hence, the main objective of the research was to utilize the classic MLP classifier algorithm combined with “sliding window” techniques capable of dynamically altering internal functions and updating over time. This approach aimed to address the mushroom classification problem, alongside additional user-friendly website offering comprehensive information to users about the content and operational procedures of the mushroom type prediction methodology.

In essence, this study integrated advanced Artificial Intelligence techniques, particularly the MLP classifier, to create a comprehensive approach for classifying mushroom types. The research journey encompassed theoretical groundwork, algorithmic exploration, and practical application development, with the ultimate aim of providing a user-friendly tool for predicting mushroom types.

2. MATERIALS AND METHODS

From the article "Edible and Poisonous Mushrooms Classification by Machine Learning Algorithms" the researchers utilized the Mushroom dataset available in the UC Irvine machine learning repository. This dataset comprises 22 features and was chosen for the study. In the process of sourcing data for the project, numerous datasets were considered. Among them, the "Mushroom Classification" dataset (UCI Machine Learning, 2016) [4] was selected due to its comprehensive specifications and high availability, featuring 23 columns and approximately 8000 records last updated six years ago. Other datasets included the "Mushroom Edibility Classification" dataset by Devzohaib (2022) [5] with 21 attributes and 61000 entries, last updated five months ago. Additionally, the "Secondary Mushroom Dataset Data Set" by Saxena (Shruti, 2022) [6] provided hierarchical data and was updated nine months ago. The "Mushroom Attributes" dataset (Pedersen, 2023) [7] was the most recent, updated on February 10, 2023. Furthermore, the "Mushroom Classification dataset" of Jha (2020) [8], featuring a dataset from two years prior, was also considered. All of these datasets were sourced from the Kangle repository. These diverse datasets served as the foundation for the research, enabling the researchers to comprehensively analyze and classify edible and poisonous mushrooms using various machine-learning algorithms, ultimately contributing to a deeper understanding of mushroom classification and safety.

The listed datasets are all relevant to predicting mushroom types; however, not all of them meet the requirements of the current research. Some datasets have inherent issues that render them unsuitable for this project. To be eligible for use in this study, a dataset must be numeric, have specific classifications, contain multiple fields for objective results, and be the most up-to-date. Among the datasets mentioned, only the "Mushroom Attributes" dataset fulfills these criteria. This dataset is a potential candidate due to its numeric nature, well-defined classifications, comprehensive attributes for objective analysis, and recent updates.

It's important to note that this dataset might be temporary in nature, as the mushroom landscape evolves and diversifies over time. This could lead to potential changes in the classification of the dataset, indicating an environment of instability. Consequently, attempting to augment the dataset by adding more classes may not be feasible using conventional machine learning algorithms, as traditional methods do not accommodate such dynamic changes. Hence, advanced algorithms capable of handling data in dynamic environments are required for this purpose. After transforming the Mushroom Attributes dataset from its raw form into a standardized format for system utilization, adhering to data input rules is essential when using the prediction system. According to the system's regulations, the predictive labels must precede the parameters, and the parameters must be placed afterwards. Following the data normalization process, a standardized dataset was obtained, comprising a total of 8124 records. This dataset includes parameters with the following format:

1:< cap-shape>2:<cap-surface> 3:<cap-color> 4:<bruises%3F> 5:< odor> 6:<gill-

attachment> 7:< gill-spacing> 8:< gill-size> 9:< gill-color> 10:< stalk-shape> 11:< stalk-root> 12:< stalk-surface-above-ring> 13:< stalk-surface-below-ring> 14:< stalk-color-above-ring> 15:<stalk-color-below-ring> 16:<veil-type> 17:<veil-color> 18:<ring-number> 19:<ring-type> 20:<spore-print-color> 21:<population> 22:< habitat> 23:<class>

This format ensures that the input data conforms to the system's requirements, facilitating accurate and efficient predictions while maintaining consistency and adherence to the specified input structure. These indices have been validated through experts in the field. Each index will have its own parameters within the database.

- Cap shape: It refers to the shape of the mushroom cap. (1: convex, 3: flat, 4: knobbed, 5: bell, 13: conical, 14: sunken)

- Cap surface: It denotes the surface texture of the mushroom cap. (3: fibrous, 8: grooves, 2: scaly, 1: smooth)

- Cap color: This represents the color of the mushroom cap. (24: brown, 5: buff, 13: cinnamon, 8: gray, 20: green, 10: pink, 16: purple, 25: red, 7: white, 2: yellow)

- Bruises? (0: no, 1: yes)

- Odor: It indicates the odor of the mushroom. (11: almond, 12: anise, 13: creosote, 2: fishy, 3: foul, 17: musty, 24: none, 10: pungent, 1: spicy)

- Gill attachment: It refers to the attachment of the gills. (11: attached, 18: descending, 3: free, 24: notched)

- Gill spacing: It represents the spacing between gills. (13: close, 7: crowded, 18: distant)

- Gill size: It describes the size of the gills. (5: broad, 24: narrow)

- Gill color: This specifies the color of the gills. (14: black, 24: brown, 5: buff, 21: chocolate, 8: gray, 20: green, 15: orange, 10: pink, 16: purple, 25: red, 7: white, 2: yellow)

- Stalk shape: It denotes the shape of the stalk. (25: enlarging, 9: tapering)

- Stalk root: This indicates the root of the stalk. (5: bulbous, 13: club, 16: cup, 25: equal, 20: rhizomorphs, 22: rooted)

- Stalk surface above ring: The surface of the stalk above the ring. (3: fibrous, 14: silky, 1: smooth, 2: scaly)

- Stalk surface below ring: The surface of the stalk below the ring. (3: fibrous, 14: silky, 1: smooth, 2: scaly)

- Stalk color above ring: The color of the stalk above the ring. (24: brown, 5: buff, 13: cinnamon, 8: gray, 15: orange, 10: pink, 25: red, 7: white, 2: yellow)

- Stalk color below ring: The color of the stalk below the ring. (24: brown, 5: buff, 13: cinnamon, 8: gray, 15: orange, 10: pink, 25: red, 7: white, 2: yellow)

- Veil type: It indicates the type of veil covering the mushroom. (10: partial, 16: universal)

- Veil color: The color of the veil covering the mushroom. (24: brown, 15: orange, 7: white, 2: yellow)

- Ring number: The number of rings on the mushroom. (24: none, 15: one, 9: two)

- Ring type: The type of ring on the mushroom. (13: cobwebby, 25: evanescent, 3: flaring, 12: large, 24: none, 10: pendant, 1: sheathing)

- Spore print color: The color of the mushroom's spore print. (14: black, 24: brown, 5: buff, 21: chocolate, 20: green, 15: orange, 16: purple, 7: white, 2: yellow)

- Population: The abundance of the mushroom. (11: abundant, 13: clustered, 24: numerous, 1: scattered, 19: several, 2: solitary)

- Habitat: The habitat where the mushroom is found. (8: grasses, 12: leaves, 10: meadows, 16: paths, 7: urban, 18: woods)

- Class: The classification of the mushroom. (10: edible, 25: poisonous)

The dataset used in this experiment consists of two parts: the training data and the testing data. The training data includes 5686 samples (which accounts for 70% of the original data), and the testing data contains 2438 samples (which accounts for 30% of the original data). The positions of these data samples will be shuffled in each experiment. Random shuffling will be performed both before training and after training. The experiment will be conducted using a batch learning approach with a batch size of 5686. This means that the system will execute 5686 batches, where each batch contains approximately 129 data samples.

All the achieved results are based on Balanced Accuracy, a metric that can be used to evaluate the performance of a binary classifier. It is particularly useful when classes are imbalanced, meaning one of the two classes appears much more frequently than the other. Using Balanced Accuracy is much more intricate than traditional accuracy. Traditional accuracy is a straightforward calculation of the percentage of a data group based on the total available data. It initially works well, but when data is heavily skewed (e.g., 1 data point in class A, 999 data points in class B), the accuracy's correctness is compromised. To address this issue, a computation method has been devised that calculates the percentage based on true negatives, true positives, false negatives, and false positives. With these parameters in place, the formula for Balanced Accuracy can be applied to calculate the most accurate and optimal percentage. Balanced Accuracy takes into account these factors to provide a more comprehensive and reliable evaluation of a classifier's performance, especially when dealing with imbalanced datasets where traditional accuracy may not be a suitable measure.

The Balanced Accuracy formula used in these experiments is:

$$\text{Balanced accuracy} = \frac{(\text{Sensitivity} + \text{Specificity})}{2}$$

The experimental dataset consists of two parts: one for training and the other for testing. The training dataset contains 5686 rows of data, and the testing dataset contains 2438 rows of data. This dataset will vary in each experiment, but the data itself is shuffled and preserved without any changes. The model used in the experiment follows a batch data approach. This means that the model utilizes the dataset in batches. The original dataset is divided into smaller groups of 50 steps each, where each batch contains 129 data points. This batch size is chosen to strike a balance between being manageable and informative for the experimentation process.

3. RESULTS AND DISCUSSION

Currently, foundational databases are facing challenges due to their lack of adaptability over time. This issue arises because these databases are trained using classical algorithms, which involve a one-time training process and require relearning from scratch whenever new data arrives. For instance, if data set 1 is trained to

create a model, when new data set 2 arrives, data set 1 needs to be relearned from the beginning alongside data set 2 to generate a new model. However, in the modern context, real-world environments experience continuous changes in data over time. Training must be carried out continuously in real-time, necessitating the constant updating of predictive models. This underscores the importance of learning from data in dynamically evolving environments. Therefore, the experimental approach must involve continuous learning methods within non-stable environments to effectively adapt to the changing data landscape. Several methods have been applied to transform classical algorithms into Continuous Learning methods, replacing them with the sliding window approach to advance traditional machine learning techniques. The description of the Sliding Windows approach is as follows:

Considering the recent development of concepts in an ever-changing training data environment, this approach involves a temporal window of fixed size (based on time intervals or the number of data points). This method can either reclassify the "group" type (based on data selected within the temporary window) or update models if online learning allows for it. In this case, the process of "forgetting" (as mentioned earlier) is automatically managed through this learning method. Typically, this method comprises three steps: Detect concept drift by using statistical tests on different windows, If an observed change occurs, select representative and recent data to adjust the models, Update the models. The window size is predetermined by the user. The key point of these methods lies in determining the window size. Most methods use a fixed-size window that is configured for each real-world problem. This allows classical algorithms to be used in dynamic environments but lacks the characteristics of incremental learning (not reusing stored data, only using the model to improve). Therefore, the historical section of the following algorithms focuses solely on presenting Incremental Learning algorithms, which have been researched and developed in recent years. By applying artificial intelligence methods, specifically algorithms like MLP classifier combined with the Sliding Window approach, a comparative analysis of these three algorithms' results can be achieved. Utilizing a chart to compare the average outcomes of these algorithms, this method ensures the most fairness in assessing the data's credibility when comparing results across different algorithms. The experimental average results of the algorithms are depicted in the chart below (Figure 2).

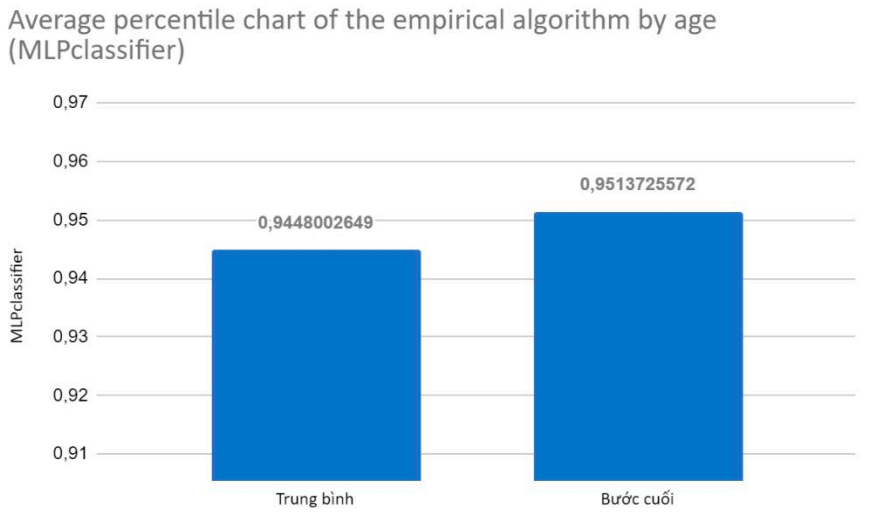

Figure 1. Average percentile chart of the empirical algorithm by age (MLP classifier)

Based on the data on the chart, we can analyze the average ratio and the final step's relatively good ratio of MLP classifier as follows:

The performance of MLP classifier is quite stable, consistently achieving above 91%: The data indicates that MLP classifier maintains stable and noteworthy performance with an average accuracy of over 94% (precisely 94.480%), and the accuracy in the final stage is also above 95% (precisely 95.137%). This is a significant advantage, demonstrating that MLP classifier has the capability to provide accurate predictions in most cases, making it suitable for addressing this stress prediction problem. In addition to calculating the algorithm's average results, another approach, such as analyzing the experimental model results by age, provides a more comprehensive and detailed view. This helps us assess the results visually and arrive at the most accurate conclusions. The results of the experimental model by age are represented in the chart below (Figure 3). This additional analysis based on age allows us to gain insights into how the model's performance varies across different age groups. It provides a more nuanced understanding of the model's strengths and weaknesses and helps in making informed decisions about its applicability and performance in different scenarios.

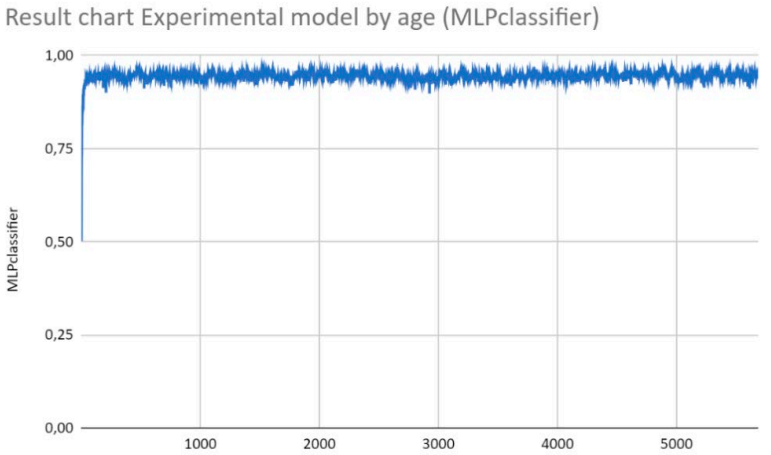

Figure 2. Result chart Experimental model by age (MLP classifier)

Considering the chart, we observe that the MLP classifier algorithm starts with a relatively low point of around 50.196% and gradually increases in the first 5 to 10 steps before stabilizing. Analyzing the chart, it's evident that MLP classifier exhibits relative stability. However, the highest accuracy achieved by MLP classifier can reach up to 96.614%, which is quite high and can be compared favorably with many other algorithms. Specifically, in the range of steps from 443 to 525, as shown in the chart below (Figure 3).

This specific analysis within the range of steps from 443 to 525 highlights the algorithm's remarkable performance during this phase. The fact that MLP classifier achieves such a high accuracy rate is an encouraging result and underscores its effectiveness in handling the given problem. It demonstrates the potential of MLP classifier as a powerful tool for accurate predictions, particularly in the specified range of steps.

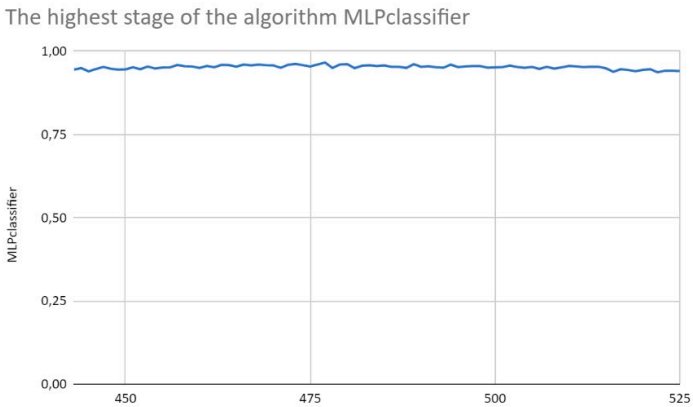

Figure 3. The highest stage of the algorithm MLP classifier (Batch 443-525)

From step 443 to step 525, the MLP classifier model demonstrated consistent and gradual performance improvement, achieving a notable accuracy rate consistently above 96.614%. This highlights MLP classifier's capability to handle non-continuous and missing value data effectively. This feature reduces the need for extensive data preprocessing, enabling the algorithm to operate efficiently across various types of data. MLP classifier consistently proves itself as a robust algorithm, even during periods where the results might not be optimal. It quickly stabilizes and maintains its performance, as depicted in the chart below (Figure 4, Figure 5). This behavior further reinforces the adaptability and reliability of MLP classifier, making it a valuable tool for various scenarios where data might exhibit variations or missing values. Its ability to maintain stability and achieve high accuracy rates, as showcased in the chart, is a strong testament to MLP classifier's effectiveness and potential for real-world applications.

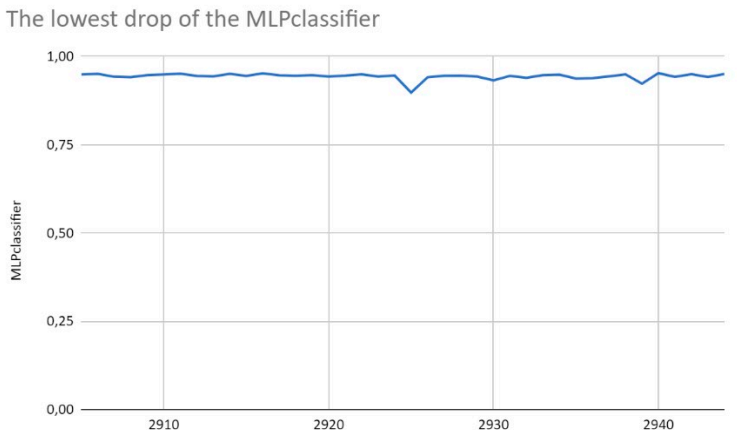

Figure 4. The lowest drop of the MLP classifier (Batch 2905 -2944)

During the MLP classifier's training phase, there was a decline in performance from step 2905 to step 2944, reaching a lowest point of 89.751% accuracy at step 2925. This could be attributed to the model struggling to learn from the data or encountering undesired variations. However, the model swiftly recovered afterward and began to improve its performance. From step 2906 to step 2920, the MLP classifier exhibited a relatively stable period following the performance dip. The accuracy rates during this phase didn't differ significantly, ranging from 94.44% to 95.34%. This indicates that the model maintained a relatively stable state.

In summary, the MLP classifier algorithm offers several advantages and maintains consistent performance with accuracy above 95%. This demonstrates that the MLP classifier can be a promising and valuable choice for various prediction and classification tasks. However, it's important to note that each algorithm has its own strengths and weaknesses, and the choice of algorithm depends on the specific requirements of the task and the characteristics of the data.

Installation:

Based on the final results presented in the previous section, the chosen algorithm to address the problem is the MLP classifier. The project will encompass various functional nodes, including prediction functionality, running classical algorithms, a list of processed models, system configuration, and user authentication. This project will be implemented within a website environment, divided into two main user roles: the Algorithm Installer (administrator or developer) and the Diagnostician (end user). These roles are illustrated in the use case diagram below:

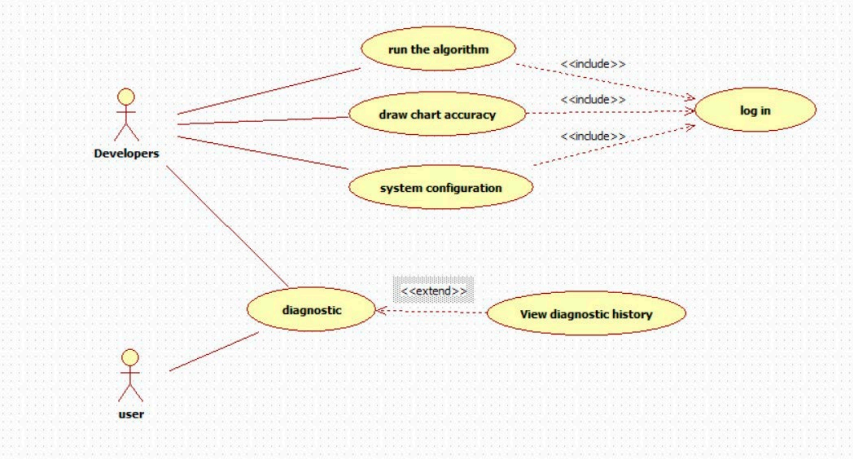

Figure 5. System use-case diagram

In order to run the software, the user's machine needs to have some Python libraries and Python version 3.9.9 installed. After extracting the contents, you will find a folder named "_SETUP_". Inside this folder, there will be a file named "python-3.9.9-amd64.exe" used to install Python 3.9.9, and a file named "inLib.bat" to install the necessary libraries required to run the software. Once the environment setup is completed, there will be a file named "Remove.bat" which is used to delete unnecessary files, including test files. This should only be used in two cases: right after extraction and installation, or to delete all previous test data. Finally, to run the program, you need to use the "Runserver.bat" file to launch the application. This file is configured to execute the command "py manage.py runserver", and the program will run on the default port "http://127.0.0.1:8000/". It's important to note that the user's computer must be connected to the internet at all times, and the minimum system requirements include Windows 10, 2GB RAM, and 10GB or more of available hard drive space to ensure smooth and stable performance. To use the software after successful installation, you can access the portal at "http://127.0.0.1:8000" to enter the main page of the system. On the main interface page, you will see a set of input fields used for making predictions. Below are the forms that have been built within the "Mushroom Prediction" system:

Figure 6. The main interface of the Mushroom Prediction system

4. CONCLUSION

Developing algorithms with multiple training options provides flexibility and high customization for the system. This allows users to experiment and choose the most suitable algorithm for specific tasks, leading to higher performance and reliability in prediction and classification. Developing a web interface in this direction makes the system user-friendly and convenient. The combination of machine learning and a website interface enables non-experts to interact easily with the system, input data, and receive classification results in a visual and clear manner. Addressing the issue of data variability by developing training algorithms in a model-based approach is a significant step in handling data fluctuations. The ability to update the model with new data helps the system maintain high and reliable performance over time. The potential for expansion and development of the topic is outlined up to the research phase, but important future directions have been proposed. This demonstrates the potential for further development and real-world application of the system. The suggested extensions include automating raw data processing, optimizing the model training process, improving the user interface, and deploying the application in practical settings. Using the MLP classifier algorithm based on the requirements for dynamic data is highly reasonable. These algorithms effectively handle changing data and help the system achieve optimal performance in various scenarios.

REFERENCES

[1] Li, H. (2021). Reviewing the world's edible mushroom species: A new evidence-based classification system, *Comprehensive Reviews in Food Science and Food Safety.* https://ift.onlinelibrary.wiley.com/doi/10.1111/1541-4337.12708.

[2] Nguyễn Thị Thu Hà (2016). *Giới thiệu về nấm.* Công ty TNHH Xuất Nhập Khẩu 2 Lúa, https://www.2lua.vn/article/gioi-thieu-ve-nam-y-nghia-va-vai-tro-cua-nam-5844d8b6e4951903508b4568.html

[3] Tutuncu, K. (2022). Edible and Poisonous Mushrooms Classification by Machine Learning Algorithms. *Mediterranean Conference on Embedded Computing (MECO),* https://ieeexplore.ieee.org/abstract/document/9797212

[4] UCI Machine Learning (2016). *Mushroom Classification.* www.kaggle.com/datasets/uciml/mushroom-classification

[5] Devzohaib (2022). *Mushroom Edibility Classification*. www.kaggle.com/datasets/devzohaib/mushroom-edibility-classification

[6] Shruti, S. (2022). *Secondary Mushroom Dataset Data Set*. www.kaggle.com/datasets/shrutisaxena/secondary-mushroom-dataset-data-set

[7] Pedersen, U.T. (2023). *Mushroom Attributes*, www.kaggle.com/datasets/ulrikthygepedersen/mushroom-attributes

[8] Jha, A.K. (2020). *Mushroom Classification Dataset.* www.kaggle.com/datasets/alokkumarjha/mushroom-classification-dataset

Journal of Science and Development Economics
Nam Can Tho University

Website: jsde.nctu.edu.vn

Performance of milk quality diagnostics using extra tree classifier techniques with progressive learning

Pham Hoang Minh[1], Truong Hung Chen[1], Pham Huynh Thuy An[1], Ngo Ho Anh Khoi[1]
[1]*Faculty of Information Technology, Nam Can Tho University*

Received: November 10, 2023
Revised: November 30, 2023
Accepted: December 5, 2023

Keywords: *AI, diagnosis, extra tree classifier, milk quality, machine learning prediction*

Từ khoá: *bộ phân loại cây bổ sung, chẩn đoán, chất lượng sữa, dự đoán, học máy, trí tuệ nhân tạo*

ABSTRACT

Quality control of milk involves the use of established control measures and testing methods to ensure proper adherence to standards and regulations concerning milk and its products. Testing ensures that dairy products meet the requirements of standards, are acceptable in terms of nutritional content, and adhere to safety standards regarding microbiological factors, heavy metals, pesticide residues, veterinary drug residues, toxins, and more. Therefore, quality checks at various stages of the milk processing chain, from farms to processing facilities and consumers, are crucial. The research method involved scientific experimentation, conducted using the Extra Tree Classifier algorithm with evolving method. The scope of the study was not extensive, and the dataset was the Milk Quality Prediction dataset sourced from kaggle.com. The aim of the study was to aid in diagnosing milk quality rapidly and relatively reliably through provided numerical data. This endeavor aims to reduce the prevalence of low-quality milk trading, ultimately contributing to safer and higher quality milk management for consumers.

TÓM TẮT

Kiểm soát chất lượng của sữa liên quan đến việc sử dụng các biện pháp kiểm soát và phương pháp kiểm tra được thiết lập để đảm bảo tuân thủ đúng các tiêu chuẩn và quy định liên quan đến sữa và các sản phẩm của nó. Thử nghiệm đảm bảo rằng các sản phẩm sữa đáp ứng các yêu cầu về tiêu chuẩn, có thể chấp nhận được về hàm lượng dinh dưỡng và tuân thủ các tiêu chuẩn an toàn liên quan đến các yếu tố vi sinh, kim loại nặng, dư lượng thuốc trừ sâu, dư lượng thuốc thú y, độc tố,... Do đó, kiểm tra

chất lượng ở các giai đoạn khác nhau của chuỗi chế biến sữa, từ các trang trại đến các cơ sở chế biến và người tiêu dùng là rất quan trọng. Phương pháp nghiên cứu này bao gồm thử nghiệm khoa học, được thực hiện bằng cách sử dụng thuật toán phân loại cây thêm với phương pháp phát triển. Phạm vi của nghiên cứu không rộng và bộ dữ liệu được sử dụng là bộ dữ liệu dự đoán chất lượng sữa có nguồn gốc từ Kaggle.com. Mục đích của nghiên cứu là hỗ trợ chẩn đoán chất lượng sữa nhanh chóng và tương đối đáng tin cậy thông qua dữ liệu được cung cấp. Nỗ lực này nhằm mục đích giảm tỷ lệ giao dịch sữa chất lượng thấp, cuối cùng góp phần vào việc quản lý sữa an toàn và chất lượng cao hơn cho người tiêu dùng.

1. INTRODUCTION

Milk quality is a crucial factor in milk production to ensure the safety of dairy products and their suitability for various purposes. To achieve quality, hygiene requirements must be implemented throughout the entire milk processing chain. Currently, with the global and specifically Vietnamese technological advancements, the counterfeiting of essential products has become an intricate issue. Dairy products, including milk, are not exceptions to this trend. Such counterfeiting can have detrimental effects on consumer health. Managing and assuring milk quality before it reaches consumers' hands is an immensely important and necessary matter, particularly in Vietnam. Milk is an essential commodity, especially for the elderly and children. It plays a vital role in enhancing health and supporting human development. Hence, if counterfeited, it can significantly impact well-being. Therefore, diagnosing and assessing milk quality hold great significance.

Elwell (2006) [1] discussed the method of using microfiltration to enhance milk quality. In this study, various milk parameters were highlighted as essential factors determining quality and extending the shelf life of milk. Marth (1981) [2], presented the interconnectedness of milk production and supply, emphasizing that ensuring high-quality milk and its derived products requires collaborative efforts from various stakeholders. The process involves producers, processors, distributors, and ultimately consumers. In another article, Marth (1972) [3] discussed the standards for quality testing of dairy products, shedding light on the criteria used to evaluate milk quality. Tường Vi (2019) [4] presentd stringent regulations to ensure milk quality before it reaches consumers' hands. In VIRAC (2017) [5], the author provided an overview of quality and food safety management, underscoring the importance of ensuring milk quality during the milk processing.

Additionally, many articles about the testing milk problem has been proposed. In Viện Kiểm Nghiệm An Toàn Vệ Sinh Thực Phẩm Quốc Gia (2022) [6], the author discussed the necessity of milk and dairy product testing, outlines regulations related to such testing, and explores the process of testing milk and dairy

products at the National Institute for Food Safety Testing. Many of them works with the AI methods. Padmaja (2021) [7] utilized the Extra Tree Classifier algorithm along with various other algorithms. Specifically, the article proposed the use of machine learning classification algorithms to predict Vitamin D Deficiency (VDD) severity. These machine learning algorithms included Random Forest (RF), Multi-Layer Perceptron (MLP), K-Nearest Neighbor (KNN), Support Vector Machine (SVM), Decision Tree (DT), Gradient Boosting (GB), Stochastic Gradient Descent (SGD), AdaBoost (AB), Extra Tree Classifier (ET), and Logistic Regression (LR). The article evaluated the output of different machine learning classification methods to estimate the severity of VDD in individuals. In the first two decades of the 21st century, a series of analyses, including Free Volatile Carboxylic Acids (FVCAs), attempted to describe 10 different types of cheese from Switzerland. The current work aimed to investigate whether these 10 types of cheese can be classified using supervised machine learning techniques and to analyze the significance of FVCAs features in understanding the role they play in describing the various types of cheese. Particularly, emphasis was placed on the SHAP (Shapley Additive exPlanations) values. In total, 241 cheese samples were classified using various ML algorithms with the assistance of the PyCaret library, with at least 90% accurately classified by two ensemble algorithms: Extra Tree and Random Forest (Bachmann, 2023) [8]. Mujumdara (2019) [9] compared multiple algorithms including Decision Tree, Gaussian NB, LDA, SVC, Random Forest, ExtraTrees, AdaBoost, Perceptron, Logistic Regression, Gradient Boost Classifier, Bagging, and KNN applied to the Diabetes Prediction dataset. This dataset consists of 800 rows and 10 fields: Number of Pregnancies, Glucose Level, Blood Pressure, Skin Thickness (mm), Insulin, BMI, Age, JobType (Office-work/Fieldwork/Machine-work), Outcome. The process includes several steps: (1) import necessary libraries, (2) load the dataset, (3) adjust paths in the training dataset, (4) compare accuracy, and (5) predict and determine the most accurate model on the test data. The employed approach can serve as a reference regarding the operational aspects of problem-solving.

2. MATERIALS AND METHODS

During the data search process for the topic, there were not a considerable number of datasets related to milk quality (approximately 135 relevant datasets). However, among them, three datasets with complete and high-quality information stood out: First, "Milk Quality Prediction" dataset by Rajendran (2022): This dataset comprises 1059 rows and 7 fields, including pH (ranging from 3 to 9.5), Temperature (ranging from 34°C to 90°C), Taste (with only 2 values: 0 and 1), Odor (with only 2 values: 0 and 1), Fat (with only 2 values: 0 and 1), Turbidity (with only 2 values: 0 and 1), and Color (ranging from 240 to 255). There is also a label field with 3 values (high, medium, low). Second, "Milk quality" dataset by Yrohit also consisted of 1059 rows and 7 fields similar to the previous dataset, with similar values, and includes a label field with 3 values (high, medium, low). Finally, "Milk Quality Prediction" dataset by Harini (2023): This dataset has a similar structure with 1059 rows and 7 fields, similar to the above datasets, along with a label field with 3 values (high, medium, low). All these datasets are related to predicting

milk quality. However, only one dataset meets the requirements of the research, including numerical data, clear classification, and multiple fields to obtain objective results. Among them, the "Milk Quality Prediction" dataset by Rajendran was selected for the research as it was updated before the other two datasets and had been verified by the scientific research community for accuracy.

The Milk Quality Prediction dataset by Shrijiayan Rajendran was last updated on 02/08/2022. The normalized dataset consists of 8 features, totaling 1060 instances, including the following parameters in the given format: <Predicted Label> 1:<pH> 2:<Temperature> 3:<Taste> 4:<Odor> 5:<Fat> 6:<Turbidity> 7:<Colour>. These indices are derived entirely from the data source mentioned above and have been validated by experts in this field. Each feature will have the following indices:

- pH: Represents the pH level, should be entered for prediction within the range of 0 – 10, with the lowest value being 3 and the highest being 9.5.

- Temperature: Describes the temperature of the milk, should be entered for prediction within the range of 0 – 100, with the lowest value being 34 and the highest being 90.

- Taste: Indicates the taste of the milk, should be entered for prediction as 0 (poor) or 1 (good).

- Odor: Indicates the odor of the milk, should be entered for prediction as 0 (poor) or 1 (good).

- Fat: Represents the fat content of the milk, should be entered for prediction as 0 (low) or 1 (high).

- Turbidity: Indicates the turbidity of the milk, should be entered for prediction as 0 (low) or 1 (high).

- Colour: Describes the color of the milk, should be entered for prediction within the range of 240 – 255, with the lowest value being 240 and the highest being 255.

The downloaded dataset can be referred to as raw data, which cannot be directly used for this research due to the need for extensive transformation from various parameters to expressed characteristics. Raw data must undergo preprocessing before it can be applied to the training process of the software. From this raw data, unnecessary information such as serial numbers and IDs will be removed. These details are not essential for the training process of the program. Next, the labels will be transformed: labels like "high," representing high milk quality, will be converted to the number "1"; labels like "medium," representing "average" quality, will be converted to the number "2"; and the remaining labels, denoting "low" quality, will be transformed into the number "3." The remaining attributes like pH (pH level), Temperature, Taste, Odor, Fat, Turbidity, and Colour will be kept unchanged as they are already in numerical form and do not require conversion. All of these conversion processes will be carried out entirely using Microsoft Excel software.

After transforming the raw data into standardized format for use in the software, the dataset will have 7 main fields: pH, Temperature, Taste, Odor, Fat, Turbidity, and Colour, along with 1 label field containing the converted values of 1 (high), 2 (medium), or 3 (low), and a total of 1,060 rows. The dataset must be saved in a file with the ".csv" extension. Each data point should be entered in a single cell, and the parameters as well as the label must be separated by a comma ",". As per the

system's requirement, the data's parameters should be placed before the label, and the label should be placed at the end. The data will be split into training and testing sets with a ratio of 70% and 30% respectively. After splitting the data, it will be saved to the computer. Next, the dataset will be split into batches for running the algorithms. The batch size will be set to 70% * n or 70% * n/2 with a 70% train/test ratio, where n represents the total number of data points.

Currently, fundamental databases are facing a significant challenge: their ability to adapt to the evolving nature of data over time. This issue arises because classical algorithms, on which current databases are often trained, can only undergo training once and need to be retrained from scratch when data changes occur. For instance, when a database is trained to create a model, with the arrival of new data, the database must start anew to generate a new model capable of accurate predictions. However, in modern data environments, data changes continuously over time, necessitating the ability to adapt rapidly to these changes.

The essence of "Concept Drift" involves the evolution and adaptation of models in response to changing environments. In non-stationary settings, the sliding window approach becomes pivotal, ensuring that models stay updated while handling variations effectively. This is especially important in domains where changes can happen rapidly or unpredictably, requiring agile methodologies to adapt and maintain performance over time.

The term "Concept Drift" has been widely used in many real-problem with the food testing. Indeed, the concept of drift forms the basis for both gradual and continuous changes and the "forgetting" of previous situations. However, challenges in non-stationary environments arise from the fact that circumstances can change rapidly or slowly, be forgotten, and even instances where knowledge reappears after it has disappeared. In such complex cases, the situation of "dual dilemmas" regarding stability or adaptation holds true. It's important to note that these approaches are not universally "incremental approaches". In the fields, three types of proposed methods are listed as solutions to this issue:

Sliding Window Approach: Considering the evolution of concepts in non-stationary environments can be accomplished by utilizing a "sliding window" method, such as the FLORA approach. This principle involves updating the model at each time point using the most recent training data, defined by a window of time or a number of data instances. This approach can perform re-classification within a "group" type (on data selected by the temporal window) or update models if online learning methods allow. In this case, the "forgetting" (as mentioned earlier) is automatically managed by this learning method. This method generally involves three steps: 1) detecting concept changes by using statistical tests on different windows; 2) if an observed change occurs, selecting representative and recent data to adjust models; 3) updating models. The window size is predetermined by the user. The key point of these methods is to determine the window size. Most methods employ a fixed-size window that is configured for each practical scenario.

Dataset Temporariness: The dataset is considered temporary due to the emergence of many new standards, resulting in increased diversity that may lead to future changes in the

dataset's classification. Consequently, this dataset can be classified as belonging to an unstable environment. Therefore, traditional machine learning algorithms used in static environments cannot be applied because conventional methods do not allow data augmentation. Instead, advanced algorithms capable of handling data in dynamic environments must be employed.

To address this issue, the concept of continuous learning in non-stable environments has been proposed as an alternative solution. Continuous learning allows databases to learn from data continuously in changing data environments, facilitating the update and adjustment of their predictive models. This enables databases to adapt to data changes and enhance the accuracy of their predictive models. Based on these characteristics and drawing from the history of research on various methods, the upcoming problem will utilize the sliding window approach. The sliding window method is the most suitable in this case, and it will be combined with the Extra Tree Classifier algorithm.

This approach enables databases to continuously adapt and update their models as new data arrives, ensuring that the predictive models remain accurate and relevant even as the data environment changes. This is particularly important in scenarios where data shifts are frequent and unpredictable, making it crucial for databases to maintain their performance and adaptability over time.

3. RESULTS AND DISCUSSION

The Sliding Windows approach, is a method commonly used to address the issue of concept drift in non-stationary environments. The key idea is to adapt the model continuously to the changing nature of the data over time. This method is often applied using techniques like FLORA. The principle involves updating the model at each time step using the most recent training data within a defined time window or a certain number of data points. This approach can involve reclassifying the "group" (based on the temporarily selected data) or updating the model if online learning techniques are applicable. In this case, the process of "forgetting" (as mentioned above) is managed automatically through this learning method. The Sliding Windows approach typically includes three main steps:

- Detecting Concept Changes: This is done by employing statistical tests on different sliding windows to identify changes in the concept.

- Adapting to Changes: If a concept change is detected, representative and recent data points are selected to adjust the models accordingly.

- Updating Models: The models are then updated based on the chosen data points to account for the concept drift.

The size of the sliding window is determined by the user and is a key parameter of these methods. Most approaches use a fixed-size window that is configured according to the specific characteristics of the real-world problem. The Sliding Windows approach is effective in capturing changes in the underlying data distribution and allows the model to adapt to these changes over time. This is particularly valuable in scenarios where the data is not stationary and concepts evolve. The ability to dynamically adjust the model's parameters and structure based on the most recent data helps maintain the model's relevance and predictive accuracy in a changing environment.

All achieved results will be based on Balanced Accuracy, which is a measure that can be used when evaluating the performance of a binary classifier. It is particularly useful when classes are imbalanced, meaning one class appears much more frequently than the other. Using Balanced Accuracy is much more complex than traditional accuracy. Regular Accuracy simply calculates the percentage of correctly classified data points based on the total available data. While this initially works well, when data is heavily skewed, the accuracy's reliability diminishes. To address this, a calculation method was devised based on true negative and true positive rates, which can compute true negative rate, true positive rate, false negative rate, and false positive rate. After obtaining these values, the Balanced Accuracy formula can be used to compute the most accurate and optimized percentage. The formula for Balanced Accuracy used in these experiments is as follows:

- First need to consult the matrix table: Absolutely, referencing a confusion matrix is a crucial step when evaluating classification results. A confusion matrix provides a clear overview of the performance of a classification algorithm by displaying the true positive, true negative, false positive, and false negative counts for each class. It's a helpful tool to understand how well your model is performing, especially in situations where classes might be imbalanced.

- Must have:

TPR (true positive rate) = TP/(TP+FN)

TNR (true negative rate) = TN/(TN+FP)

- After obtaining the above two parameters, Balanced Accuracy is calculated using the formula:

Balance Accuracy = (TPR+TNR)/2

Based on the data presented in the chart, we can observe the following average ratios and the final step of the Extra Tree Classifier algorithm:

The performance of the Extra Tree Classifier appears to be quite stable, with an average ratio across all steps reaching 60%, specifically 70.12%. The average accuracy rate of the final step also reaches 75%. This indicates that the Extra Tree Classifier is capable of making reasonably accurate predictions in various scenarios, with relatively good accuracy. Therefore, the Extra Tree Classifier algorithm seems suitable for addressing the milk quality prediction problem, as it demonstrates the ability to provide fairly accurate predictions with a relatively high level of accuracy (Figure 2).

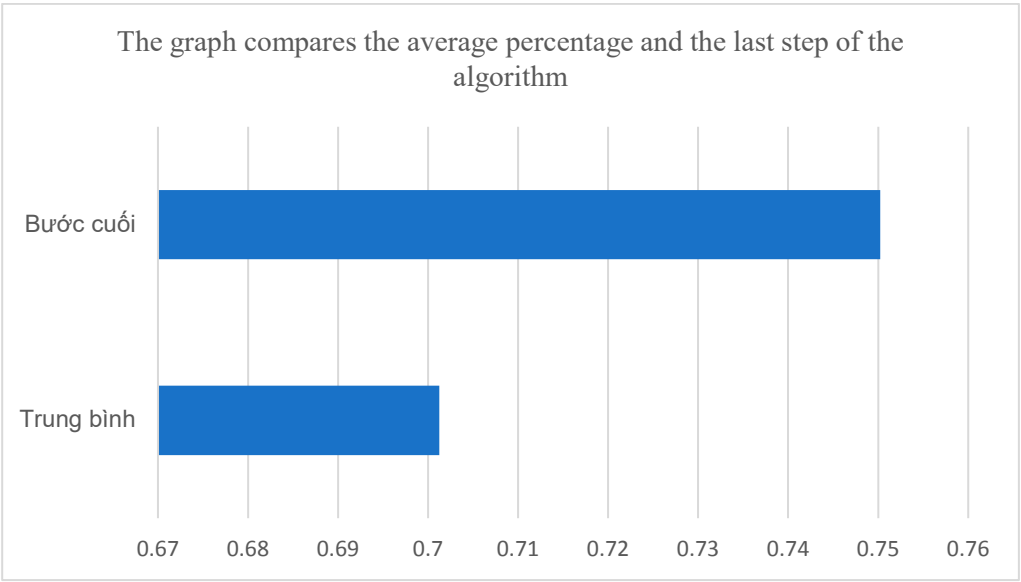

Figure 1. The graph compares the average percentage and the last step of the algorithm (Extra Tree Classifier)

Sure, you can compare the performance of the Extra Tree Classifier algorithm in more detail by analyzing the accuracy progression for each step. The accuracy progression chart of the Extra Tree Classifier algorithm provides a clearer and more comprehensive view of its performance. This chart can help you evaluate the algorithm's behavior and effectiveness throughout its execution.

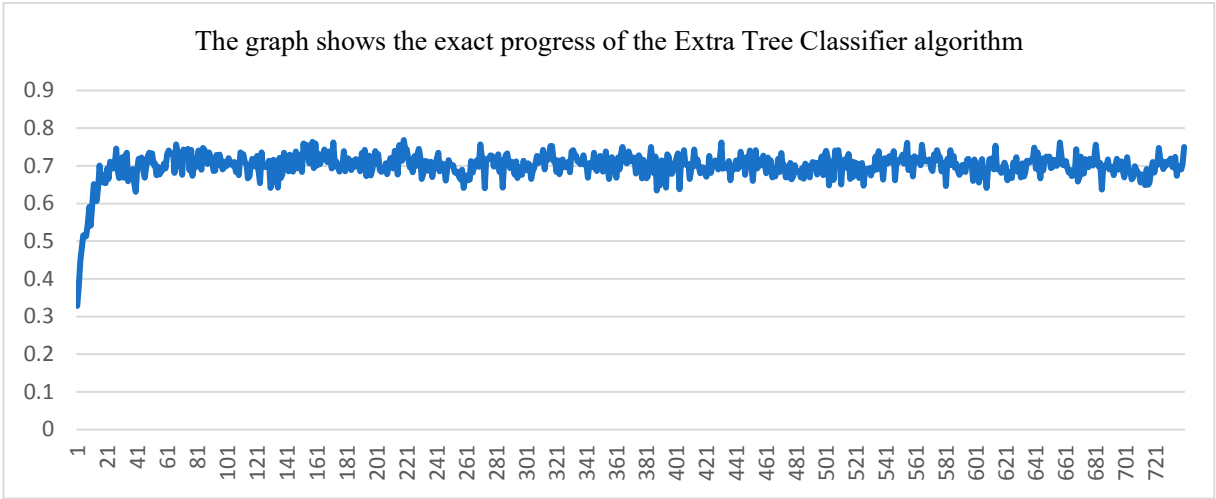

Figure 2. The graph shows the exact progress of the Extra Tree Classifier algorithm

Based on your description of the accuracy progression chart for the Extra Tree Classifier algorithm, it seems that the algorithm starts with a relatively low accuracy of 37% but gradually improves its stability. The most noticeable improvement occurs from step 1 to step 20. In

the beginning, the algorithm's average accuracy is only 31%, but by step 20, it increases to 70%, and then reaches a stable level. The accuracy of the Extra Tree Classifier seems to show relatively stable fluctuations ranging from 60% to a peak of 76.22%. The highest accuracy is achieved at step 219, with a rate of 76.86%. This is a substantial value that can be compared favorably to many other algorithms. Other steps with high average accuracy include step 172 (76.21%), step 270 (75.67%), step 318 (75.3%), step 431 (76.18%), and step 555 (76.12%). Furthermore, you mention that the most significant growth in accuracy occurs between steps 421 and 431. This information provides valuable insights into how the algorithm performs and where its strengths lie.

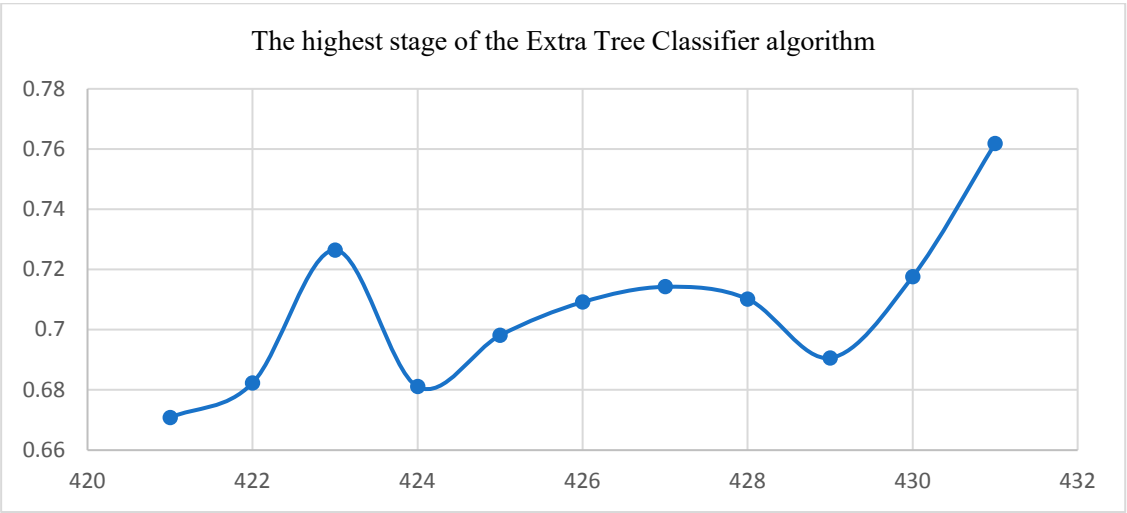

Figure 3. The highest stage of the Extra Tree Classifier algorithm (Batch 421-431)

Observing the chart, it is evident that there is stability and gradual improvement in the accuracy progression of the Extra Tree Classifier algorithm. The accuracy consistently reaches above 60% by the step with the highest rate. This level of stability is quite satisfactory compared to other algorithms. Despite occasional periods of less favorable results during data processing, the Extra Tree Classifier consistently stabilizes relatively quickly, as depicted in the chart below:

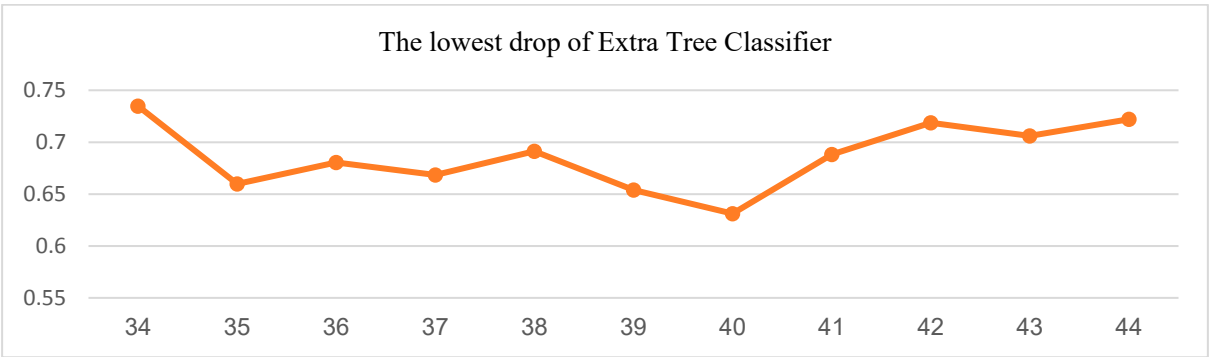

Figure 4. The lowest drop of Extra Tree Classifier (Batch 34 -44)

At step 40, the Extra Tree Classifier achieves its lowest accuracy rate of 63.11%. However, by step 44, the algorithm quickly regains stability with an accuracy of 72.2%, and it gradually stabilizes in the subsequent steps. Although this period represents the lowest accuracy phase for the Extra Tree Classifier, the accuracy remains above 60%, indicating a reasonably stable phase. The discrepancies in accuracy values are not too large during this phase. For example, in the range from step 35 (65.96%) to step 39 (65.4%), there is a slight drop in accuracy from step 34 (73.47%) to step 35 (65.96%), followed by stability from step 35 to step 38 (69.13%), a deep decrease to 63.11% at step 40, and then an increase back to 72.2% at step 44.

From the comparative charts, it can be observed that the Extra Tree Classifier algorithm demonstrates relatively good stability with limited fluctuations in the average accuracy rates across steps. The accuracy of the algorithm is also notably high, ranging from 60% to 80%. However, it's important to note that each algorithm has its own strengths and weaknesses. Choosing an appropriate algorithm depends on the specific requirements of the problem and the characteristics of the dataset at hand.

Application installation:

Based on the final results presented earlier, the Extra Tree Classifier algorithm will be chosen as the method to address the issue of milk quality diagnosis in the application. The system has been completed and meets all the initial requirements, and it is divided into distinct sections to serve both regular users and developers. Regular users will have an interface to perform milk quality diagnosis. On the other hand, developers will have more interface pages to work with and develop within the system. These interface pages include the login page, data training page, diagnostic function testing page, model list display page, and settings page. The system has a high practical applicability, especially in the context of increasing milk consumption in Vietnam. Over time, regulations regarding milk quality may change due to various objective reasons. Therefore, the dataset will also undergo significant changes. With the Extra Tree Classifier algorithm, the system is capable of effectively addressing and adapting to these changes.

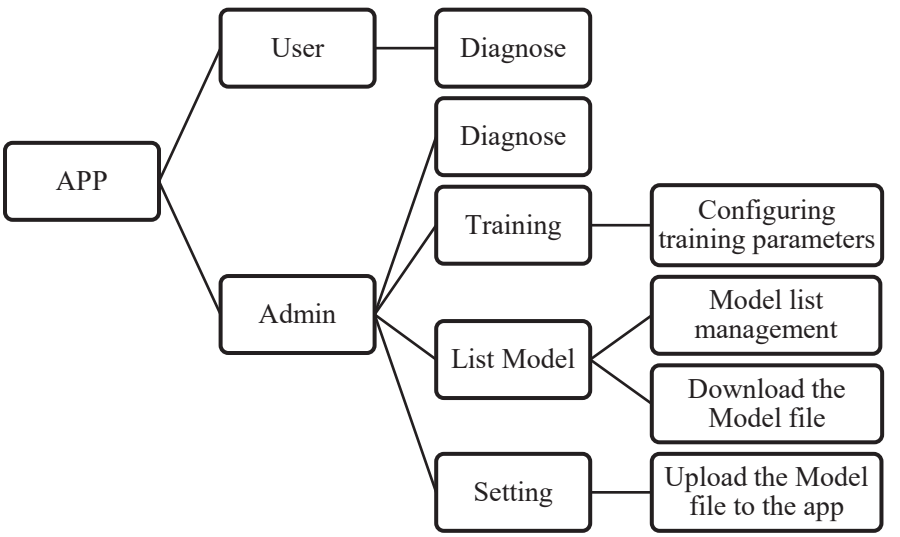

Figure 5. Use case diagram of the system

The section for predictors:

Figure 6. Main interface of Milk Quality Diagnostic system

The main interface of the app is where users can perform the diagnosis. Users need to input complete information into various input fields. The required information includes the 7 attributes of milk quality: pH level, temperature, taste, turbidity, fat content, odor, and color. After providing these inputs, users can press the "Start Prediction" button to receive the prediction results.

In this project, the system has been packaged as a .ZIP file named "ChanDoanSua-main.zip". After users download and extract it, they will find a folder named "BaoCaoThucTap". Inside this folder, there are files to run the program. Below are the installation instructions:

The system requires a computer that is always connected to the internet and meets the minimum configuration: Windows 10, 2GB RAM, and 10GB hard drive space. To run the software, follow these steps:

- Go to the "SETUP" folder and install Python. The version used is Python 3.9.9 (python-3.9.9-amd64.exe).

- Install the required libraries by running the file "CaiThuVien.bat".

- Run the "RunServer.bat" file to start the program.

- Open a web browser and navigate to the address 'http://127.0.0.1:8000/'.

- The program will run on the default port: "http://127.0.0.1:8000/".

The program can also be deployed on a web platform. This is the default installation process.

4. CONCLUSION

The project has achieved several significant outcomes. Firstly, it involved a comprehensive study of milk quality trends both globally and in Vietnam, providing insights into the crucial role of quality assessment. The project successfully managed the database, selecting and transforming it to meet the project's specific needs. Additionally, the project delved into the workings and practicality of the Extra Tree Classifier algorithm, effectively utilizing it to address the problem at hand. An essential aspect of the project was the meticulous comparison of

algorithm stability, resulting in the identification of the most suitable approach. This culminated in the development of a diagnostic application featuring an intuitive and user-friendly interface. Looking ahead, the project envisions further enhancements. This includes refining the user interface for better aesthetics and usability, expanding the application's capability to diagnose diverse datasets from various domains beyond milk quality, and refining the algorithm for heightened accuracy. The project's context is rooted in the Vietnamese milk market, acknowledging concerns about subpar milk quality affecting both export potential and public health. By leveraging the Extra Tree Classifier algorithm alongside sliding window techniques, the system adapts to dynamic data changes. The software, developed as a web-based application, ensures an accessible interface that caters to developers and end-users alike. The visual system model encapsulates user interaction and developer installation procedures, using commonly employed software tools.

REFERENCES

[1] Elwell (2006*). Use of microfiltration to improve fluid milk quality.* https://www.sciencedirect.com/science/article/pii/S002203020672361X

[2] Marth (1981). *Assuring the quality of milk.* https://www.sciencedirect.com/science/article/pii/S0022030281826793

[3] Marth (1972). *The thirteenth edition of standard methods for the examination of dairy products.* https://www.sciencedirect.com/science/article/pii/S0362028X23022202

[4] Tường Vi (2019). *7 nguyên tắc vàng mang đến từng hộp sữa tươi an toàn và chất lượng.* https://vov.vn/suc-khoe/7-nguyen-tac-vang-mang-den-tung-hop-sua-tuoi-an-toan-va-chat-luong-964919.vov

[5] VIRAC. (2017). *Tổng quan ngành công nghiệp sữa.* https://investvietnam.gov.vn/vi/nghanh.nghd/15/sua-va-cac-san-pham-sua.html

[6] Viện Kiểm Nghiệm An Toàn Vệ Sinh Thực Phẩm Quốc Gia (2022). *Kiểm nghiệm sữa và sản phẩm sữa.* https://nifc.gov.vn/ky-thuat-chuyen-mon/kiem-nghiem-sua-va-san-pham-sua-post1484.html

[7] Padmaja, B. (2021). *Prognosis of Vitamin D deficiency severity using SMOTE optimized Machine Learning Models*. https://turcomat.org/index.php/turkbilmat/article/view/8442

[8] Bachmann, H.P. (2023). *Classification of cheese varieties from Switzerland using machine learning methods: Free volatile carboxylic acids.* https://www.sciencedirect.com/science/article/pii/S0023643823006746.

[9] Mujumdara, A. (2019). *Diabetes Prediction using Machine Learning Algorithms.* https://www.sciencedirect.com/science/article/pii/S1877050920300557

Journal of Science and Development Economics
Nam Can Tho University

Website: jsde.nctu.edu.vn

Water quality prediction using MLP in dynamic environment

Tran Van An[1], Kieu Tien Binh[1], Nguyen Dinh Thuy Huong[2], Ngo Ho Anh Khoi[1]
[1]*Faculty of Information Technology, Nam Can Tho University*
[2]*Vietnam Maritime University*

Received: November 10, 2023
Revised: November 30, 2023
Accepted: December 5, 2023

Keyword: *mssmartyPants, MLP classifier, water quality*

Từ khoá: *chất lượng nước, phân loại MLP*

ABSTRACT

Water quality plays an immensely significant role in safeguarding the health of humans and other species on the planet. Polluted water sources can contain harmful substances such as heavy metals, pesticides, bacteria, and viruses, which can pose threats to human health and other organisms. Water quality is also a crucial factor in protecting the environment and sustaining the ecosystems of freshwater and marine environments. Polluted water sources can impact the lives of aquatic animals and plants, causing harm and endangering biodiversity by disrupting the food chain. Within the scope of this article, we utilize data from "Water Quality" by MssmartyPants using the classical MLP Classifier algorithm. Incorporating artificial intelligence into water quality assessment contributes to increased accuracy and efficiency compared to manual methods, while also providing significant value in academic research.

TÓM TẮT

Chất lượng nước đóng một vai trò vô cùng quan trọng trong việc bảo vệ sức khỏe của con người và các loài khác trên hành tinh. Nguồn nước bị ô nhiễm có thể chứa các chất có hại như kim loại nặng, thuốc trừ sâu, vi khuẩn và vi rút, có thể gây nguy hiểm cho sức khỏe con người và các sinh vật khác. Chất lượng nước cũng là yếu tố quan trọng trong việc bảo vệ môi trường và duy trì hệ sinh thái nước ngọt và môi trường biển. Nguồn nước bị ô nhiễm có thể ảnh hưởng đến đời sống của động vật và thực vật thủy sinh, gây hại và gây nguy hiểm cho đa dạng sinh học do làm gián đoạn chuỗi thức ăn. Trong phạm vi bài viết này, chúng tôi sử dụng dữ liệu từ "Chất lượng nước" của MssmartyPants bằng

thuật toán Phân loại MLP cổ điển. Việc kết hợp trí tuệ nhân tạo vào đánh giá chất lượng nước góp phần tăng độ chính xác và hiệu quả so với phương pháp thủ công, đồng thời mang lại giá trị đáng kể trong nghiên cứu học thuật.

1. INTRODUCTION

The issue of clean water is a global concern and is regarded as one of the major challenges for sustainable development worldwide. According to reports from the World Health Organization (WHO) and UNICEF, approximately 2.2 billion people worldwide still lack access to clean water, and 4.2 billion people lack access to sanitation services (UNESCO, 2016) [1]. Vietnam is among the countries facing the challenge of clean water. Despite having significant water sources like the Red River, Mekong River, and the Gulf of Tonkin, water quality in many regions is encountering difficulties, affecting both human health and the environment. Water sources are polluted due to various factors including waste, livestock waste, chemical-intensive agriculture, industrial activities, and rampant littering by the population - as reported by the Ministry of Natural Resources and Environment, around 60-70% of lakes, river, (Ministry of Natural Resources and Environment, 2023).

UNESCO (2016) [1] discussed how water quality is one of the key challenges that society will face in the 21st century, posing threats to human health, limiting food production, diminishing the ecological functions of ecosystems, and impeding economic growth. In the article "Drinking-water - World Health Organization (WHO)," the author emphasizes the importance of clean and easily accessible water for public health and highlights how improving water supply and sanitation can drive economic growth for countries and significantly contribute to poverty reduction (WHO, 2022) [2]. Lastly, Roy et al. (2019) [3], highlighted how water can be one of the most valuable natural resources after air. In Vietnam, we have several articles addressing water quality issues. Huỳnh Phú et al., (2021) [4] focused on analyzing surface water quality based on the economic development of regions in Bac Lieu Province. Next, Vũ Thị Thanh Hương et al. (2020) [5] predicted water quality in the Bac Hung Hai irrigation system according to socioeconomic development scenarios up to the year 2020. Lastly, Vũ Thị Hồng Nghĩa et al. (2011) [6] presented evaluations of the water quality of the Cau River and proposed environmental management solutions to protect and improve water quality.

Lê Phước Cường et al. (2020) [7] utilized machine learning models such as Linear Regression, Random Forest, Support Vector Machine, K-nearest neighbor, and Cubist to predict groundwater quality near the Cẩm Hà landfill, Hoi An City. Rosly et al. (2015) [8] compared various classification methods such as Naive Bayes (NB), J48 Decision Tree, Sequential Minimal Optimization (SMO), Multi-Layer Perceptron (MLP), and Instance-Based Learning with k-Nearest Neighbors (IBK) for water quality classification of the Kinta River dataset in Perak, Malaysia. Results showed that multi-classification approaches could achieve higher accuracy than individual methods [8]. Nasir et al. (2022) [9] discussed

the use of seven individual classifiers to predict the Water Quality Index (WQI). The stacked model proved successful in predicting water quality, and the CATBOOST method yielded the best prediction results.

Finally, while there are no specific mentions of water quality prediction using the MLP Classifier algorithm, similar algorithms have been used to introduce innovative approaches. This algorithm is likely to yield the best results for water quality diagnosis. The dataset will be sourced from Kangle and the selected dataset named "Water Quality" by MssmartyPants.

2. MATERIALS AND METHODS

Most of the data used for research and learning purposes in this article will come from Kaggle, including the "Water Quality" dataset by Aditya Kadiwal. This dataset contains a fair amount of data errors (missing or lost data in certain rows/columns), which can significantly impact subsequent machine-learning efforts (Kadiwal, 2021) [10]. The "Indian water quality data" by Anbarivan and Anjali Vasudevan contained text-based information, making data normalization challenging, and it's not suitable for classification algorithms, thus not fitting the current purpose of the article of Ramakrishnan (2016) [11]. Lastly, the "Water Quality" dataset by MssmartyPants, comprising 21 columns and over 8000 rows of data, is relatively well-structured and has been extensively used for scientific research purposes, making it the chosen dataset (MsSmartyPants, 2021) [12].

Dataset:

All the listed datasets are related to water quality prediction; however, only one dataset meets the requirements of the research topic, as the others have various issues when used in this context. The dataset must be numerical, have specific classifications, and contain multiple fields to yield objective results. Only the Water Quality dataset from MssmartyPants meets these criteria, which is why it was selected for the internship project. However, this dataset is temporary in nature, as the indices can differ between databases, potentially leading to changes in the dataset's classification in the future. This highlights the fact that this dataset exists in an unstable environment, making it unsuitable for using classical machine learning algorithms in a static environment to increase dataset class sizes. Instead, advanced algorithms need to be employed to handle data in a dynamic environment. Raw data must be processed before being applied to the software's training process. The attributes used for testing, such as aluminum, ammonia, arsenic, barium, cadmium, chloramine, chromium, copper, fluoride, bacteria, viruses, lead, nitrates, nitrites, mercury, perchlorate, radium, selenium, silver, uranium, will be kept unchanged as they are numerical values, requiring no conversion. All conversion processes are entirely conducted within Microsoft Excel software.

After transforming the data from its raw form to a standardized format for software use, the dataset must be saved as a ".csv" file extension. Only one cell is allowed for each data entry, and parameters as well as labels must be separated by commas. According to system regulations, data parameters must come first, followed by the label. After all data standardization processes, a standardized dataset containing 21 features with a total of 8000 data points will be obtained. These indices are sourced from the aforementioned data and have been verified by experts in the field. Each feature will have the following indices:

Prediction label: This parameter carries a decisive value and is particularly important. The label can only take one of two values: "0" for unsafe or "1" for safe. The predictions consist of a total of 7084 unsafe data points and 916 safe data points.

Aluminum: It is found everywhere, in food, water, and cooking utensils made of aluminum. Scientists have found that aluminum has negative effects on health. High doses of aluminum exposure can damage bones and tissues. Bones lose calcium and phosphorus, becoming weaker and causing bone pain (ranges within the dataset from 0 to 5.05).

Ammonia: A colorless gas with a pungent odor, chemical formula NH_3. Ammonia is not highly toxic to humans and animals. However, if it exceeds permissible levels in water, it can transform into cancer-causing agents and other dangerous diseases (ranges within the dataset from 0 to 29.84).

Arsenic: Also known as arsenic, a highly toxic compound, four times more toxic than mercury. Ingesting water with even half a grain of rice's worth of arsenic can kill a healthy person. The World Health Organization (WHO) has reported that for every 10,000 cancer cases, 6 deaths are attributed to water with arsenic levels above the standard of 0.01 mg/l (ranges within the dataset from 0 to 1.05).

Barium: Barium is a solid substance that contributes to pollution in various wastewater treatment systems today. However, most people still have little understanding of the hidden potential dangers and optimal treatment methods for this pollutant in wastewater, tap water, and industrial water (ranges from 0 to 4.94 in the dataset).

Cadmium: Cadmium is one of the three most dangerous metals for the human body, the other two being lead and mercury. Regular consumption of water containing cadmium can increase the risk of diseases such as prostate cancer and lung cancer (ranges from 0 to 0.13 in the dataset).

Chloramine: Not only does chloramine have an unpleasant odor, but it also affects your health. The impact depends on the chlorine residue in the water. According to QCVN 01:2009/ BYT (National Technical Regulation on Drinking Water Quality), the permissible chlorine level in water is 0.3-0.5 mg/l. However, the actual situation currently shows a widespread excess of chlorine (above the standard) in tap water (ranges from 0 to 8.68 in the dataset).

Chromium: Chromium is a heavy metal that can cause cancer if accumulated in the body. According to the World Health Organization, chromium is toxic to the body, and drinking water containing chromium, even in amounts as low as 1-2g, can lead to immediate death (ranges from 0 to 0.9 in the dataset).

Copper: Copper is a fairly common metal found in water. To ensure safety for users, the copper content in water must be less than 2mg/l. The harmful effects of this heavy metal in water include irritation and corrosion of mucous membranes, nerve inhibition, etc (ranges from 0 to 2 in the dataset).

Fluoride: Accumulation of excessive fluoride in the body can affect joints, leading to increased risks of joint pain, immobility, weakened bones, and even bone cancer. Additionally, excessive fluoride in water can increase the risk of thyroid gland diseases (ranges from 0 to 1.5 in the dataset).

Bacteria: The presence of hidden microorganisms in water indicates that the water source is not safe, causing various dangerous symptoms such as diarrhea, vomiting, high fever, etc (ranges from 0 to 1 in the dataset).

Viruses: Similar to bacteria, the presence of viruses in water is extremely dangerous (ranges from 0 to 1 in the dataset).

Lead: The presence of lead in water is due to the corrosion of pipes and industrial wastewater. According to current regulations on clean water and drinking water, the lead content in water must not exceed 0.01 mg/l (ranges from 0 to 0.2 in the dataset).

Nitrates: If water with nitrate is heated, it can form nitrosamines. There are various types of nitrosamines, some of which can increase the risk of cancer (ranges from 0 to 19.83 in the dataset).

Nitrites: Similar to nitrates, water with nitrites heated at high temperatures is extremely dangerous (ranges from 0 to 2.93 in the dataset).

Mercury: Mercury is a metallic element found naturally in air, water, and soil. Even slight exposure to mercury can cause severe health problems, threatening the development of fetuses and the early stages of children's lives. Apart from young children, mercury poisoning can harm the nervous, digestive, and immune systems, affecting the lungs, kidneys, skin, and eyes (ranges from 0 to 0.01 in the dataset).

Perchlorate: Formed from one chlorine atom and four oxygen atoms, perchlorate is a powerful oxidant widely used in rocket fuel, fireworks, and flares. They are also produced as byproducts from chemical manufacturing and herbicides, and therefore, after use, they can contaminate the environment, and seep into water sources, and soil (ranges from 0 to 60.01 in the dataset).

Radium: Health issues related to radium stem from its radioactive properties. Radium primarily appears in groundwater in the form of radioactive isotopes radium-226 and radium-228. As these isotopes decay, they emit alpha particles that can damage human tissue. Alpha radiation is blocked by the skin, so there is no danger when bathing or washing dishes with water containing radium. However, long-term consumption of water with radium can lead to chronic health problems, including an increased risk of bone cancer (ranges from 0 to 7.99 in the dataset).

Selenium: While selenium is recognized as an essential element for human health, it is only needed in very small amounts (the human body requires very little selenium). Prolonged exposure above the permissible limit can lead to selenium poisoning, with symptoms such as depression, anxiety, nervousness, mood swings, nausea, vomiting, garlic-scented breath and sweat, and in some cases, hair loss and brittle nails (ranges from 0 to 0.1 in the dataset).

Silver: When nano silver particles come into contact with the human body, they attack the immune system, destroy cell structures, and gradually weaken health. This prolonged condition can increase the risk of devastating diseases such as Alzheimer's, Parkinson's, and even cancer (ranges from 0 to 0.5 in the dataset).

Uranium: In nature, uranium (VI) forms highly soluble carbonate complexes in alkaline environments. This enhances the mobility and persistence of uranium in soil and groundwater, originating from nuclear waste materials, posing health risks to humans (ranges from 0 to 0.09 in the dataset).

Overall, this database has been relatively well standardized, except the author not providing additional information about certain data fields, requiring further investigation. The experimental dataset consists of two parts: a training dataset containing 5600 data points (70% of the original data) and a test dataset containing 2400 data points (30% of the original data). The positions of these data points will change in each experiment, and each experiment will involve random shuffling both during training and afterward.

The experiment will be conducted using a batch learning model with a batch size of 699. This means the system will perform 699 steps, with each step containing about 129 data points. The model used in the experiment is a batch data model. This model will use the same batch data set by dividing the original data set into smaller batches in 35 steps, meaning each batch contains 129 data points. This number is neither too large nor too small, making it convenient for experimentation.

3. RESULTS AND DISCUSSION

The sliding window approach is a technique used in machine learning to train models by utilizing small data windows. These windows are created by sequentially moving through the dataset, generating a new window at each step. The size of the window can be fixed or variable, and the overlap between consecutive windows can also be controlled. This approach is computationally efficient and can be employed for performing machine learning tasks on data streams. This training model method has several advantages compared to other methods, such as batch learning. Firstly, it enables faster training times as it processes only a small portion of data at each step. Secondly, it can help avoid overload as the model continuously interacts with new data points. Lastly, it is memory-efficient, requiring only a small portion of data to be loaded into memory at a time. Therefore, employing this approach significantly enhances the effectiveness and productivity of ML algorithms (Joseph, 2022) [13].

The model used for conducting scientific experiments has been explicitly discussed in the previous section. Therefore, in this section, the focus is on analyzing and comparing results among various algorithms.

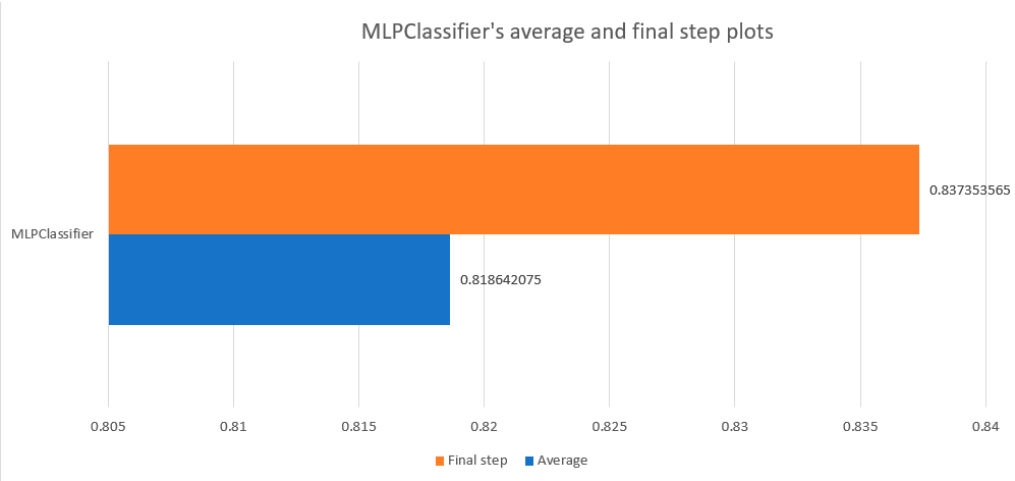

Figure 1. Average of BA and in final step

The MLP Classifier is a machine learning algorithm used for classification tasks. The average curve and final accuracy are typically important metrics for evaluating the performance of a classification model. In your specific case, the average curve accuracy is 0.818, and the final accuracy is 0.837. Accuracy is computed by comparing the number of correct predictions to the total number of samples in the test dataset. The average curve is a graph illustrating how the model's accuracy changes as different parameters or conditions are altered. Through the average curve, you can identify certain parameter thresholds that lead to improved model performance, aiding in optimizing its effectiveness. The final accuracy, on the other hand, represents the ultimate accuracy of the model after optimization and training on the training data. This is the accuracy the model can achieve when used in real-world applications. The difference between the average curve and the final accuracy reflects the model's optimism or realism. If the average curve exhibits higher accuracy than the final accuracy, it could suggest that further fine-tuning of the model could enhance its performance. Conversely, if the final accuracy closely resembles the average curve, this indicates that the model has been well-optimized and can be directly applied to new data. In summary, the average curve and final accuracy of the MLP Classifier provide important information about the classification model's performance on test data. Analyzing and evaluating these metrics helps you gain a better understanding of the model's capabilities and limitations in classifying different samples.

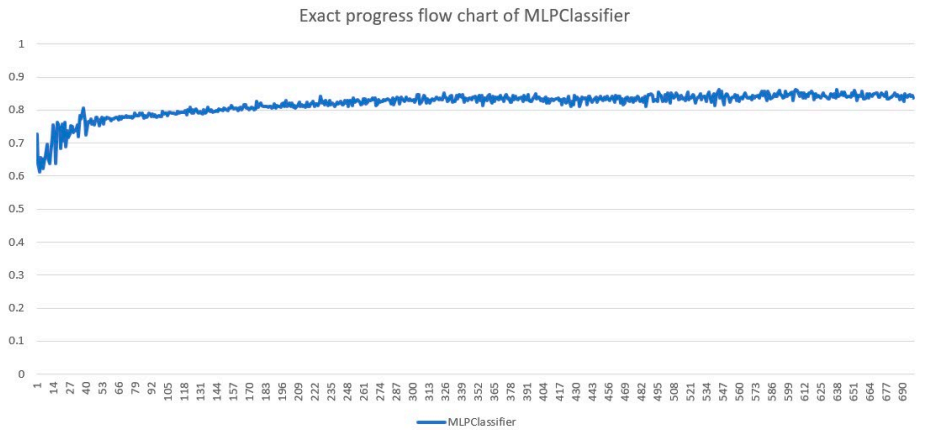

Figure 2. Progress flow chart of MLP Classifier

Increasing Variation Trend: The accuracy progression plot of the MLP Classifier illustrates that the model's accuracy increases over time or the number of iterations. This indicates that the model is learning from the data and improving its accuracy as time goes on. The gradual increase in accuracy suggests that the model is not just initially randomly learning from the data but also capturing more important features as training progresses further. This could imply that the model is enhancing its performance by grasping the data's complexity better. The fluctuation within the range of 0.814 to 0.855 indicates that the model's accuracy is

not stable and varies across iterations. However, this variation falls within a narrow range, which may suggest that the model is nearing the threshold of maximum performance for the current training dataset.

In summary, the accuracy progression plot of the MLP Classifier depicts a model evolving from an initial accuracy of 0.814 to a higher level of 0.855 overtime or iterations, despite minor fluctuations during this process+.

Installation:

MLP Classifier will be the chosen algorithm for the web environment. It includes three forms: the login form, the diagnosis form, and the settings form. To use the software after successful installation, you need to access the portal "http://127.0.0.1:8000" to enter the main page of the system. Here, the login interface will be displayed. In the login interface, there are two ways to log into the system: one is to log in as a developer and the other is to log in as a user.

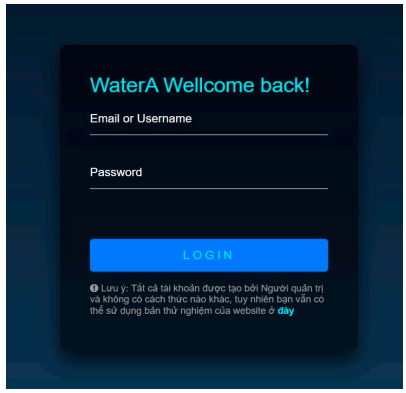

Figure 3. Login screen

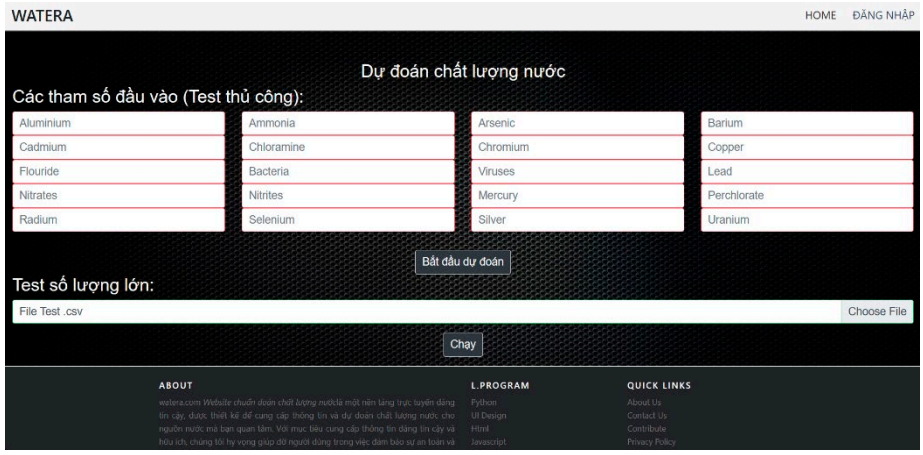

Figure 4. Main interface

As the project is still under development, user account allocation remains limited. Therefore, apart from using the provided accounts, users can also access the website directly to experience its main functionalities. There are two methods for data processing: manual data entry or processing a large amount of data using a ".csv" file. Here is the main interface:

Interface for the list of trained models:

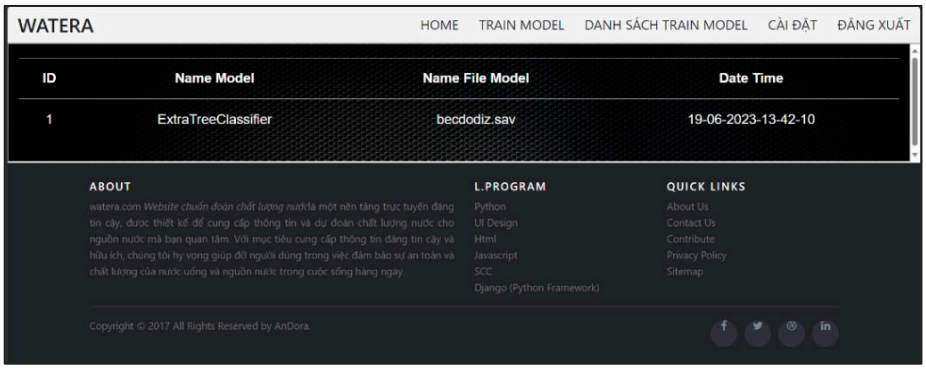

Figure 5. Model installation interface

Configuration settings interface:

System Requirements: Operating System: Windows 10, RAM: Minimum 2GB, Hard Drive Space: Minimum 10GB, and Internet Connection.

Installation process:

Step 1: Install Python and Libraries: Extract the files and open the "SETUP" folder, run the "python-3.9.9-amd64.exe" file to install Python 3.9.9, after Python installation is complete, run the "inLib.bat" file to install the necessary libraries.

Step 2: Remove Excess Data (Optional):

If needed, you can run the "Remove.bat" file to delete unnecessary data files. Only perform this step in the mentioned cases.

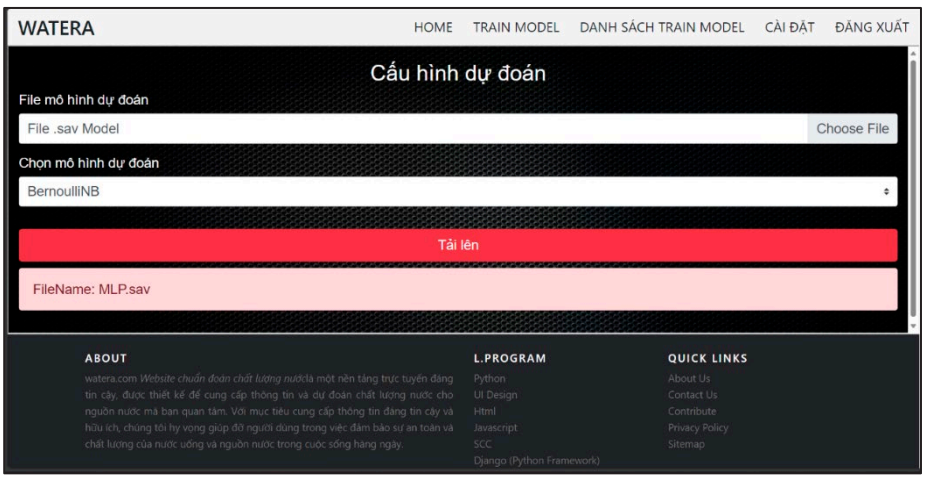

Figure 6. Configuration settings interface

Step 3: Run the Program: To run the program, use the "Runserver.bat" file. This file is configured to execute the command "py manage.py runserver". The program will run on the default port "http://127.0.0.1:8000/".

Using the Software: Ensure your computer is always connected to the internet. Access the address "http://127.0.0.1:8000" to reach the main page of the system. On the main interface page, you will see a template of input fields to predict water quality.

Note that I have used the information you provided to create a general guide. If you encounter any issues or have specific questions about the installation, usage, or operation of the software, you should refer to the documentation provided by the developer or the software's support team.

4. CONCLUSION AND SUGGESTIONS

In conclusion, we have explored three classical algorithms used in the project, delved into the specifics of the "Water Quality" database by MssmartyPants, and explained the need for a specific development direction to effectively utilize the application. This research enables us to comprehend the significance of water quality and the integration of artificial intelligence into diagnostic processes. Water quality remains a pressing concern in our nation's developmental journey, as the lack of clean water sources can impact socioeconomic factors, facilitate disease transmission, and affect human health. Ultimately, prevention is better than cure, and together, we must foster a collective awareness to safeguard clean water sources and mitigate actions that contribute to environmental pollution, especially in water sources.

REFERENCES

[1] UNESCO. (2016). *The global water quality challenge & SDGs.* https://en.unesco.org/waterquality-iiwq/wq-challenge

[2] WHO. (2022). *Drinking-water*. https://www.who.int/news-room/fact-sheets/detail/drinking-water. Retrieved May 13, 2023.

[3] Roy, R. (2019). *An Introduction to Water Quality Analysis.* https://www.researchgate.net/publication/352907194

[4] Huỳnh Phú (2021). *Nghiên cứu phân vùng chất lượng nước mặt theo diễn biến phát triển các vùng kinh tế của tỉnh Bạc Liêu.* https://www.researchgate.net/publication/352250212_Nghien_cuu_phan_vung_chat_luong_nuoc_mat_theo_dien_bien_phat_trien_cac_vung_kinh_te_cua_tinh_Bac_Lieu

[5] Vũ Thị Thanh Hương, Nguyễn Đức Phong, & Nguyễn Xuân Khôi (2020). Nghiên cứu dự báo chất lượng nước trong hệ thống thủy lợi bắc Hưng Hải theo các kịch bản phát triển kinh tế xã hội đến năm 2020. *Tạp Chí Khoa Học và Công Nghệ Thủy Lợi*

[6] Vũ Thị Hồng Nghĩa (2011). *Đánh giá hiện trạng chất lượng nước sông Cầu và đề xuất giải pháp quản lý môi trường nước sông Cầu trên địa bàn Tỉnh Thái Nguyên.* Nghiên cứu quản lý chất lượng nước Sông Cầu trên địa bàn Tỉnh Thái Nguyên (vnu.edu.vn)

[7] Lê, P., & Cường (2020). *Ứng dụng mô hình học máy dự báo chất lượng nước dưới đất: Điển hình tại khu vực thành phố Hội An, Tỉnh Quảng Nam (Application of Machine Learning Models in underground water prediction: A case study in Hoian City, Quangnam Province.* https://media.neliti.com/media/publications/453597-application-of-machine-learning-models-i-ddd67fa8.pdf

[8] Rosly, R., Makhtar, M., Awang, M.K., & Deris, M.M. (2015). *Multi-Classifier models to improve the accuracy of water quality application.*

https://www.researchgate.net/publication/331208114

[9] Nasir, N., Kansal, A., Alshaltone, O., Barneih, F., Sameer, M., Shanableh, A., & Al-Shamma'a, A. (2022). Water quality classification using machine learning algorithms. *Journal of Water Process Engineering*, *48*, 102920. https://doi.org/10.1016/j.jwpe.2022.102920

[10] Kadiwal, A. (2021). *Water Quality*. https://www.kaggle.com/datasets/adityakadiwal/water-potability, Retrieved May 13, 2023.

[11] Ramakrishnan, V. (2016). *India water quality data.* https://www.kaggle.com/datasets/venkatramakrishnan/india-water-quality-data. Retrieved May 13, 2023.

[12] MsSmartyPants. (2021). *Water quality*. https://www.kaggle.com/datasets/mssmartypants/water-quality. Retrieved May 13, 2023.

[13] Joseph (2022). *Sliding window machine learning: What you need to know - reason*. https://reason.town/sliding-window-machine-learning/. Retrieved June 3, 2023.

Journal of Science and Development Economics
Nam Can Tho University

Website: jsde.nctu.edu.vn

Knowledge management in the 21st century: trends, developments, and strategies

Tong Wooi, CHOW (Jerry)[1*]

[1]*Malaysia University of Science and Technology, Malaysia*

[*]*Corresponding author: CHOW (Jerry) (email: jerrychow@must.edu.my)*

Received: November 10, 2023
Revised: November 30, 2023
Accepted: December 5, 2023

Keywords: *21st-century, case studies, knowledge management, strategies, trend and development*

Từ khoá: *chiến lược, nghiên cứu trường hợp, quản lý tri thức, Thế kỷ 21, xu hướng và phát triển*

ABSTRACT

In the dynamic landscape of the 21 st-century, the realm of knowledge management has gone through huge transformations, shaped by means of the relentless tempo of technological developments and evolving organizational paradigms. This article delved into the modern trends, developments, and techniques that outline the exercise of expertise administration in the digital era. In this era of information abundance, organizations are increasingly recognizing the pivotal role of effective knowledge management in achieving sustainable success. With the advent of digital technology, there has been a surge in the volume of information being produced which has led to the thriving need for effective knowledge management (KM) practices. Furthermore, this article explored the innovative developments that have re-shaped expertise knowledge management practices. It gave insights into the strategic approaches and best practices employed with the aid of forward-thinking businesses to harness the full potential of knowledge management in the digital age. From fostering a culture of knowledge sharing to ensuring strategies that are essential for navigating the complexities of present-day data ecosystems. As the knowledge management landscape continues to evolve, this article served as a treasured useful resource for professionals, researchers, and corporations searching for to adapt and thrive in a generation described through the non-stop pursuit of understanding excellence and innovation. It underscoreds the significance of embracing these trends, developments, and strategies to remain competitive and resilient in the ever-changing global

commercial environment. This article included the following firms as case studies: Xerox Corporation, Siemens AG, Carnegie Mellon University's School of Computer Science and Imperial College London's Data Science Institute. The paper concluded that knowledge management is required for organizational efficiency and success in the digital era and outlines insights on strategies for best practices. The author also included the framework on the recommended strategies for knowledge management. Findings affirmed that organizations need to adopt effective and efficient knowledge management practices for organizations to stay competitive and improve performance.

TÓM TẮT

Trong bối cảnh năng động của thế kỷ 21, lĩnh vực quản lý tri thức đã trải qua những biến đổi to lớn, được định hình bởi nhịp độ phát triển không ngừng của công nghệ và các mô hình tổ chức đang phát triển. Bài viết này đi sâu vào các xu hướng, sự phát triển và kỹ thuật hiện đại nhằm phác thảo việc thực hiện quản trị chuyên môn trong kỷ nguyên số. Trong thời đại thông tin dồi dào này, các tổ chức ngày càng nhận ra vai trò then chốt của quản lý kiến thức hiệu quả trong việc đạt được thành công bền vững. Với sự ra đời của công nghệ kỹ thuật số, khối lượng thông tin được tạo ra đã tăng vọt, dẫn đến nhu cầu ngày càng tăng về các phương pháp quản lý kiến thức hiệu quả. Hơn nữa, bài viết này khám phá những phát triển đổi mới đã định hình lại các hoạt động quản lý kiến thức chuyên môn. Nó cung cấp cái nhìn sâu sắc về các phương pháp tiếp cận chiến lược và thực tiễn tốt nhất được áp dụng với sự hỗ trợ của các doanh nghiệp có tư duy tiến bộ nhằm khai thác toàn bộ tiềm năng của quản lý kiến thức trong thời đại kỹ thuật số. Từ việc thúc đẩy văn hóa chia sẻ kiến thức đến đảm bảo các chiến lược cần thiết để điều hướng sự phức tạp của hệ sinh thái dữ liệu ngày nay. Khi bối cảnh quản lý kiến thức tiếp tục phát triển, bài viết này đóng vai trò là nguồn tài nguyên hữu ích cho các chuyên gia, nhà nghiên cứu và tập đoàn đang tìm cách thích nghi và phát triển trong một thế hệ được mô tả thông qua việc không ngừng theo đuổi sự hiểu biết về sự xuất sắc và đổi mới. Nó nhấn mạnh tầm quan trọng của việc nắm bắt những xu hướng, sự phát triển và chiến lược này để duy trì tính cạnh tranh và kiên cường trong môi trường thương mại toàn cầu luôn thay đổi. Bài viết này bao gồm các công ty sau đây làm nghiên cứu điển hình:

Tập đoàn Xerox, Siemens AG, Trường Khoa học Máy tính của Đại học Carnegie Mellon và Viện Khoa học Dữ liệu của Đại học Hoàng gia Luân Đôn. Bài viết kết luận rằng quản lý kiến thức là cần thiết để đạt được hiệu quả và thành công của tổ chức trong kỷ nguyên kỹ thuật số, đồng thời đưa ra những hiểu biết sâu sắc về các chiến lược để thực hành tốt nhất. Tác giả cũng đưa ra khuôn khổ về các chiến lược được đề xuất để quản lý kiến thức. Các phát hiện khẳng định rằng các tổ chức cần áp dụng các biện pháp quản lý kiến thức hiệu quả và hiệu quả để duy trì tính cạnh tranh và cải thiện hiệu suất.

1. INTRODUCTION

In today's fast-evolving and complex environment, knowledge management has emerged to be an important factor in organizational performance. Traditional methods are inadequate for dealing with swift environmental changes. Every organization produces, manages, and uses large amounts of information daily. The digital era has resulted in a rapid growth of data and information. Hence, managing knowledge has become an essential strategy to maintain a competitive advantage for optimum performance (Idrees et al., 2023) [1]. According to Al-Shahrani (2019) [2], knowledge management is currently a highly popular subject in both industry and information research circles. Despite the prevalence of digital technology, knowledge management remains a relatively new and constantly evolving area of management. It is regarded as a major advancement in information studies and management science. Usman et al. (2020) [3] states that effective and efficient knowledge management (KM) is essential for organizations. Moreover, the digital era has opened new avenues for KM (Manesh et al., 2020) [4].

KM is a structured methodology that encompasses the approaches of creating, capturing, refining, storing, managing, and disseminating know-how with the motive of enjoyable the desires of a business enterprise as illustrated in the expertise administration cycle (Figure 1) (Girard & Girard, 2015) [5]. Bill Gates states, "knowledge management is a fancy term for a simple idea – you're managing data, documents, and people efforts" (Sharma, 2014) [6]. The management of information gives advantages via lowering the effort and price concerned in duplicating previous efforts. Collaborating and using shared information creates value. It entails regulating expertise and its software in organizational practices inside the enterprise. Therefore, it is indispensable to recognize the evolving trends, development, and undertake the first-rate techniques for high quality KM in the digital technology to enhance organizational overall performance (Alshammari et al., 2020) [7]. This paper aims to discover the evolving trend, development, and the techniques of expertise administration or KM in the digital era.

The study will focus on the trend and development of KM in the digital era, including exploring the various strategies that organizations can adopt to manage knowledge in the digital age (Toma, 2006) [8]. The paper

traced the early development trend of the field in terms of the proponents and the KM concepts to highlighting three key strategies and examples of four organizations that have implemented KM. In terms of the limitation, the paper focuses on the guiding theme of the topic on knowledge management in the 21 st-century: trends, developments, and strategies.

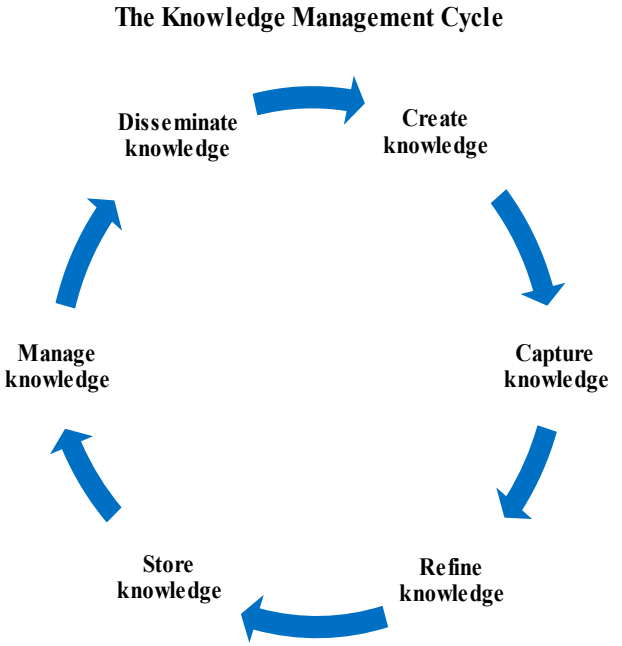

Figure 1. The Knowledge Management Cycle

2. MATERIALS AND METHODS

The narrative review was conducted using a search technique to identify and synthesize relevant publications in the following databases: Emarald Insight, Semantic Scholar, Science Direct, and Google Scholar using the keywords "knowledge management," "digital era," "KM strategy," "KM trend," and KM development." The search was limited to English language. Journal articles, conference papers, books, and edited volumes were included in the review if they met the following criteria: (1) discussed the trends, development, and strategies for effective KM in the 21 st-century; (2) case studies of organizations that has implemented KM operation (3) were published in peer-reviewed journals; and (4) were available in full-text form. The four-case studies selection was based on purposive sampling and information available in line with the topic for review. Relevant information was extracted from the selected articles, including author, year of publication, country of origin, research objectives, research approach, key findings, and recommendations for effective KM. The extracted information was synthesized into themes based on the key findings, and recommendations. The themes were then analysed to identify the evolving trend, development, and strategies for effective KM in

the digital era. The research approach is appropriate for this paper in view of the search for the information needed that traced to the early days of KM development process until the discussion on the needed strategies and relevant case studies.

2.1 The background of knowledge management

Drucker (1989) [9], a well-known expert in management, mentioned that expertise has ended up an integral financial aid and a widespread supply of aggressive advantage. As a result, know-how is an asset that corporations need to possess to reap boom and success. It is crucial for corporations to recognize the vital concepts of information and manipulate their understanding sources correctly and effectively (Roshchin et al., 2022) [10]. Knowledge administration is now not genuinely some other aid such as labor or capital; it is a fundamental useful resource that must be prioritized, as referred to by means of Ganapathy et al., (2020) [11].

The origins of KM can be traced returned to Greek philosophers such as Aristotle, who sought to create and file know-how for realistic use. The concept gained more recognition in the early 20th century, as organizations began to understand the value of knowledge as an asset. With the advancement of technology and the growth of the knowledge economy in the latter half of the 20th century, the need for effective KM became even more pressing. Drucker was one of the first to focus on information and knowledge, while Senge emphasized the concept of the *Learning Organization*, which served as a foundation for KM. By the 1980s, the importance of knowledge as a competitive advantage became increasingly apparent.

Evidence suggests that the management of knowledge depended on the utilization of artificial intelligence and expert systems. Scholarly articles, publications, and conferences started to cover topics on KM from the 1990s to present times (Mohajan, 2017) [12]. Consulting firms initiated in-house KM programs employing Adam's model. With the media coverage of KM, it gained popularity and continued to evolve as a concept, becoming essential for organizations. Consequently, companies acknowledged the significance of managing their knowledge assets and began implementing knowledge management practices.

The term *knowledge management* was coined by both Karl-Erik Svelby and Karl Wiig in 1986. Karl-Eric Sveiby pioneered many integral principles of information management. He was once described as one of the founding fathers of KM. In 1986, he published his first book *Knowledge Companies* in Sweden. On the other hand, Karl Wiig is a management researcher. He is also described as the founding father of KM. He wrote many articles and books on KM. Another influential approach in the teaching of KM is by Nonaka and Takeuchi (1995) [13], which emphasizes the importance of knowledge creation and innovation in organizations. Both Nonaka and Takeuchi essentially taught on the idea that knowledge creation is the key to organizational innovation and success (Kinyata, 2014) [14]. They stressed that organizations must create new knowledge by combining existing knowledge and expertise. They proposed a model known as the Spiral Model.

2.2 The evolving trend of knowledge management

The notion of KM is relatively recent and emphasizes the significance of managing knowledge on par with managing resources.

With the advent of the knowledge economy, a new era of management has emerged which places greater emphasis on KM. The modern organization is about knowledge. Management in the twenty-first century involves KM and is based on based on knowledge (Toma, 2006) [8]. Essentially, KM pertains to the management of knowledge within organizations and encompasses a diverse range of activities, such as generating, acquiring, organizing, and distributing knowledge (Igbinovia, 2018) [15]. In the early years, KM targeted the formation and management of explicit knowledge which essentially was done through document management systems, databases, and other related tools. However, in the advent of the digital era, KM has included the management of tacit knowledge. Digital mechanisms have enabled it easier to capture and manage this kind of knowledge. Over the years, researchers from different parts of the world have provided varying definitions of the KM discipline. Some argued that it cannot be limited to a single definition and that it is perceived differently across different fields. As a result, there seems to be no agreement on a single definition for KM.

In this paper, the definition of KM will rely on the concepts expressed by Ammirato et al. (2021) who defined KM as the comprehensive process of identifying, organizing, transferring, and utilizing information and skills. The early definition by Davenport and Prusak (1998) [16], has described KM as the process of collecting, arranging, and preserving the information and experiences of individuals and teams within an organization, and sharing it with others. According to Girard and Girard (2015) [5] and Igbinovia (2018) [15], KM seeks to help a company achieve a competitive edge by gathering these materials in a centralized or dispersed electronic setting.

To put it simply, KM involves a wide range of activities aimed at identifying, collecting, organizing, sharing, and transferring important information and expertise that make up an organization's memory. The *purpose* of a knowledge system is to maximize an organization's effectiveness and returns from its knowledge assets. The objective of knowledge management is to increase an organization's efficiency and preserve its knowledge. Early researchers such as Davenport (1998) have initially proposed four main goals of KM systems in practice: establishing knowledge repositories, enhancing knowledge access, improving the knowledge environment, and managing knowledge as an asset.

According to a study conducted by Price Waterhouse Coopers and the World Economic Forum, 95% of CEOs consider KM to be a crucial factor in a company's success. Similarly, another survey conducted among CEOs produced a comparable result regarding the importance of KM (Sardjono & Firdaus, 2020) [17]. KM can take organizations to new levels of efficiency, effectiveness, and operational reach. By enhancing operational processes, it can improve an organization's performance and financial value. KM supports sustainable strategic competitive advantage for organizations, making it an essential element for their continuous development. In short, KM has increasingly become a source of competitive advantage.

In the context of the topic, it is important to distinguish between data, information, knowledge, and wisdom, as they represent the fundamental concepts of each term (Bellinger et

al., 2004) [18]. Data refers to the raw, unprocessed elements of information in an organization. Information, on the other hand, is data that has been processed and given meaning, answering questions such as who, what, where, and when. Knowledge is the application of both data and information. Wisdom is the evaluated understanding that comes from the utilization of accumulated knowledge. In the realm of KM, three categories of knowledge are generally recognized which are explicit, implicit, and tacit knowledge. *Explicit* knowledge, also known as formal knowledge, is codified, and can be easily transformed. It is typically found in physical formats such as books, databases, memos, and electronic media that can be obtained, recorded, communicated, shared, and stored. Some examples of explicit knowledge include strategies, methods, processes, patents, products, and services.

Implicit knowledge is knowledge that builds upon existing explicit knowledge and includes transferable skills that can be applied in different jobs. Examples of implicit knowledge include data obtained from communication channels such as Skype, email, intranet, and meeting notes.

Tacit knowledge, on the other hand, is not codified and resides in individuals' minds (Nonaka & Takeuchi, 1995) [13]. This type of knowledge includes expertise, experience, skills, and technical know-how, and can be shared through mentoring, face-to-face communication, training, group projects, and other means. Tacit knowledge is not easily expressed or formalized, unlike implicit knowledge which is an application of explicit knowledge. Examples of tacit knowledge include hands-on skills, intuitions, experiences, relationships, personal beliefs and values, and ideas. Hence, organizations need to develop strategies to harness their intellectual capital (Nunes et al., 2017) [19]. Th explicit and tacit knowledge can be leveraged upon for KM best practice (Ismail & Abdullah, 2016) [20].

2.3 The development of knowledge management

KM has emerged due to various factors. The fast-paced changes in the marketplace have made it difficult for organizations to acquire knowledge and experience, leading to information overload. Additionally, organizations face pressure to reduce costs due to competition. High staff turnover has resulted in a need to develop informal knowledge using formal methods. Changes in organizational direction have also contributed to the loss of knowledge. Furthermore, life-long learning has become increasingly important.

The digital era has further transformed the way organizations manage their knowledge (Roshchin et al., 2022) [10]. The digital era has reformed the domain of KM, making it more available and adept. The progression of KM has been pushed by various factors, including globalization, technical advancements, and changing customer preferences. This has resulted in a shift from traditional KM practices to technology-driven advances. New technologies were introduced that enhance KM in organizations (Usman et al., 2020) [3].

The development of computer technology, internet, social media, cloud computing, and artificial intelligence has facilitated the process of identifying, capturing, and analyzing data and information. Mobile technology has made it possible for people to access knowledge easily, which increases efficiency. These means

have made it easier for individuals to share knowledge and collaborate on projects. This has caused the emergence of new KM systems. The proliferation of social media contributed to the culture of knowledge sharing within organizations and collaboration.

In recent times, the use of artificial intelligence and machine learning is happening rather swiftly. There are reports that artificial intelligence and machine learning are tested to automate KM processes, such as data extraction and analysis (Bughin et al., 2018) [21]. Artificial intelligence can analyze big amount of data to extract insights. Artificial learning powered KM in organizations may become a reality. Then with cloud computing, data are easily store and accessed from any place globally. This undoubtedly will facilitate the development of KM systems. At the same time, KM has become more dynamic and interactive. It is likely that there will be greater use of artificial intelligence and machine learning in KM in the future.

2.4 The strategies for knowledge management

In view of the digital era, organizations must adopt effective strategies for best practice in line with the latest trends and development for effective KM. Effective KM can facilitate better decisions, increase innovation, and employee satisfaction. Researchers were proposing for an integrative framework to support the implementation of KM (Nunes et. al., 2017) [19]. Evidence from research indicated that KM best practices utilized an integrated model approach. The strategies for KM proposed here are knowledge-sharing culture, technology, and KM strategies which are as follows (Figure 2).

2.4.1 Knowledge-sharing Culture

Developing a culture of *knowledge-sharing* is a key strategy for effective KM, as suggested in various studies. To prioritize knowledge management, organizations should establish a culture of knowledge-sharing that permeates the company. This involves encouraging employees to share their knowledge and expertise with each other, as emphasized by researchers. The knowledge-sharing culture will strengthen individual commitment and teamwork. This can be achieved through various programs and activities such as training programs, recognition, and rewards for knowledge sharing. Additionally, those at the senior management level can play a part in leading by example in promoting knowledge-sharing and emphasizing the importance of knowledge sharing. Apart from knowledge-sharing culture, the use of technology can also enhance the efficiency of KM processes (Usman, et al., 2020) [3].

2.4.2 Technology

Technology plays a crucial role in enhancing KM practices within organizations (Manesh et al., 2020) [4]. The leveraging of technology facilitates KM in the digital age. Technology can be instrumental in KM by enabling knowledge to be easily located, retrieved, and shared through KM systems. For instance, the use of KM systems (KMS) as a centralized repository can facilitate easy storage and retrieval of knowledge assets (Dalkir, 2017) [22]. Artificial intelligence and machine learning technologies can automate knowledge extraction from unstructured data, provide real-time answers through chatbots, and analyze large data sets to uncover hidden patterns and insights (Alavi & Leidner, 2001) [23]. Social

media and online collaboration tools can foster knowledge sharing and collaboration among employees (Wasko & Faraj, 2005) [24]. Data analytics and visualization tools can provide meaningful insights from knowledge assets, aiding decision-making (Alvesson & Karreman, 2011) [25]. Learning management systems (LMS) can provide access to online training programs, enabling employees to continuously upskill and acquire new knowledge (Davenport & Prusak, 2000) [16]. Technology can also facilitate innovation through idea management platforms, promoting creativity and knowledge creation (Chen, 2016) [26]. In addition, technology can improve communication among employees.

Organizations should invest in technology that supports their KM strategies. Technology facilitates KM in terms of providing tools for capturing, organizing, and sharing knowledge within the systems. Technology enhances people to communicate better. At the same time, technology improves the efficiency of knowledge management processes. Some examples of technology used for KM include management systems, social collaboration platforms and knowledge bases. The KM systems enable the organizations to organize their information in structured format making it user friendly. The systems used for KM comprise of different types such as content management systems, document management systems, and knowledge bases.

2.4.3 Knowledge Management Strategies

To achieve effective KM in the digital age, it is important to have a well-defined *system* or *process* in place for knowledge management (Hlatshwayo, 2019) [27]. Organizations need to develop a strategy that shows their goals and objectives for KM (Velazco et al., 2021). The strategy should include processes such as knowledge capturing, knowledge sharing, and knowledge dissemination. The processes must be aligned with their organizational goals and objectives and support the KM activities (Alavi & Leidners, 2001) [23]. Concurrently, the process must be adaptable for change to be effective in the dynamic digital era. Other practical suggestions include starting small. Organizations can start with a small pilot project and expand it gradually. Involve employees in the development and implementation of the KM systems. Use a variety of KM tools to capture and organize (Muhaja, 2017). The final aspect to consider in the strategies for KM is to create a plan for managing knowledge and to assess the success of the KM initiative. These procedures are crucial in ensuring that KM is efficiently implemented in organizations. There are benefits in KM in organizations. It would be appropriate to consider the challenges in implementing KM. There are various KM implementation barriers and some of them are organizational barriers, human barriers, technical barriers, financial barriers, and political barriers (Ganapathy et al., 2020) [11]. The main difficulty in managing knowledge is ensuring that the appropriate information is accessible to suitable individuals when it is needed.

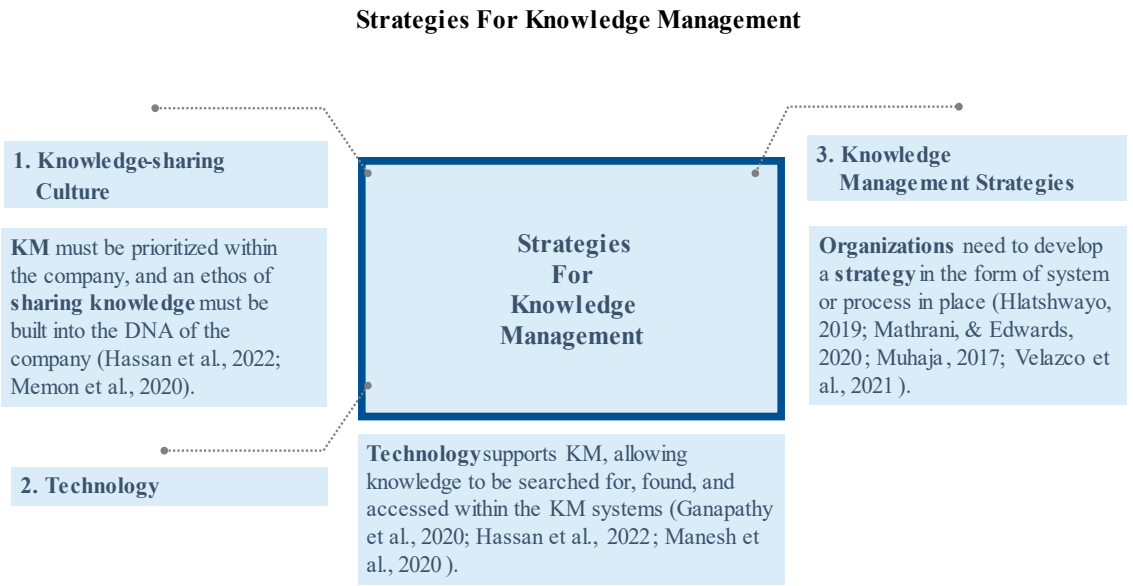

Figure 2. Strategies for Knowledge Management

3. RESULTS AND DISCUSSION

Literature has shown that KM is becoming increasingly important for organizations to achieve their business goals and stay competitive in today's rapidly changing business environment (Roshchin et al., 2022) [10]. Many renounced multinational companies, such as Xerox, Siemens, IBM, Hewlett Packard, Shell, British Petroleum, Ford, and Caterpillar, to name only a few, have implemented some forms of knowledge sharing systems. This section presents an overview of four case studies which include two leading business corporations namely, Xerox Corporation and Siemens AG, and two educational organizations, namely, Carnegie Mellon University's School of Computer Science (SCS) and the Imperial College London's Data Science Institute. These case studies are to examine their key implementation of KM and identify lessons gleaned from their experiences. These four organizations have implemented KM practices and have gained significant benefits.

3.1 Case Study 1: Xerox Corporation

The case study about Xerox Corporation presents treasured insights into the implementation of understanding KM practices in a real-world organizational context. Xerox Corporation is a main international science enterprise that specializes in file management, imaging, and associated services. Xerox has a long-standing dedication to know-how management, which has been a giant element in its success. Xerox has carried out several information administration initiatives, inclusive of the improvement of knowledge-sharing platforms, knowledge-based systems, and know-how repositories. These initiatives have helped Xerox to streamline its commercial

enterprise processes, beautify consumer satisfaction, and extend operational efficiency. One key lesson from the Xerox case study is the importance of leadership support and commitment to drive KM initiatives. Xerox's top management was committed to knowledge management and provided the necessary resources and support for its successful implementation. The management team also ensured that the knowledge management initiatives were aligned with the company's overall strategic objectives. This top-down approach was crucial in ensuring that KM was integrated into the organization's culture.

Another important lesson learned from Xerox's KM initiatives is the importance of a collaborative way of life (Powers, 1999) [28]. Xerox encouraged collaboration and knowledge sharing among its employees by creating knowledge-sharing platforms, such as the Xerox Collaborative Knowledge Exchange (CKE). The CKE provided employees with access to knowledge and expertise across the organization, enabling them to solve problems and make better decisions. Also, Xerox recognized the need for continuous learning and improvement, with regular monitoring and evaluation of KM initiatives and feedback loops for improvement (O'Dell & Grayson, 1998). Xerox's KM initiatives were periodically evaluated to identify areas of improvement and refine their KM practices. Xerox Corporation's implementation of KM has resulted in several benefits (Hickins, 2013) [29]. The company has managed to leverage its intellectual capital to develop innovative solutions that meet customers' needs. Xerox Corporation's KM strategy has also enabled the company to reduce costs and improve operational efficiency. The company has also managed to retain its employees, who feel valued and appreciated for their knowledge and expertise. Xerox Corporation's knowledge management strategy has also enabled the company to remain competitive in the global market. Xerox Corporation's profitable implementation of KM has enabled the company to leverage its intellectual capital to develop innovative solutions that meet customers' needs. The company has managed to create a culture that encourages knowledge sharing among employees, and it has invested in knowledge capture and storage technologies to ensure that critical knowledge is not lost. Xerox Corporation's KM strategy has resulted in several benefits, including improved operational efficiency, reduced costs, and a competitive advantage in the global market. The company's success in implementing KM serves as a model for other businesses that want to leverage their intellectual capital to remain competitive in the global market.

3.2 Case Study 2: Siemens AG

Siemens AG is a German multinational conglomerate that operates in several sectors such as energy, healthcare, infrastructure, transportation, and industrial automation. Siemens AG recognized the importance of KM in the late 1990s and began implementing a comprehensive KM strategy across the organization. Siemens AG has successfully implemented KM practices, which have helped the company to optimize its operations, enhance its innovation capacity, and improve its decision-making processes. This section highlights the key lessons learned from the Siemens AG case study on successful implementation of KM. One key lesson from

the Siemens AG case learn about is the significance of aligning information administration with strategic enterprise dreams and goals (Sveiby, 2001) [30]. Siemens AG ensured that know-how administration initiatives had been aligned with the average enterprise approach and supported the organization's strategic priorities. Another lesson is the significance of leadership involvement in driving KM initiatives (Alavi & Tiwana, 2002) [31].

The second lesson that Siemens AG learned was the importance of creating a knowledge sharing culture. The company recognized that KM was not just about capturing and storing knowledge, but also about making it available to employees throughout the organization. Therefore, Siemens AG applied a variety of initiatives to inspire expertise sharing, such as communities of practice, understanding sharing events, and a expertise sharing portal. These initiatives helped to create a sub-culture the place personnel had been inclined and keen to share their information with their colleagues. Additionally, Siemens AG identified the significance of technological know-how as an enabler of expertise management, via imposing a complete know-how administration device that facilitated effortless storage, retrieval, and sharing of expertise property (Riemer, 2001) [32]. However, the organization additionally identified that technological know-how used to be no longer ample and that it wanted to be supported via advantageous strategies and practices. Therefore, Siemens AG invested in coaching and improvement applications to assist personnel apprehend how to use science effectively.

Siemens AG also emphasized the significance of continuous learning and improvement, with regular monitoring and assessment of know-how KM initiatives to identify areas of improvement and refine KM practices (Alavi & Tiwana, 2002) [31]. Lastly, Siemens AG recognized the need for effective change management and communication to ensure successful adoption and integration of KM practices across the organization (Sveiby, 2001) [30]. They provided training, communication, and change management efforts to support employees in embracing and utilizing knowledge management practices. Siemens AG Corporation's implementation of KM initiatives provides valuable lessons for other organizations looking to enhance their KM practices. The successful implementation of KM initiatives at Siemens AG Corporation was attributed to the importance of establishing clear goals and objectives, creating a knowledge sharing culture, investing in the right technology, measuring the impact of its KM initiatives, and continuous improvement.

3.3 Case Study 3: Carnegie Mellon University's School of Computer Science (SCS)

KM performs a pivotal function in the success of instructional establishments and businesses alike. This case find out about explores how Carnegie Mellon University's School of Computer Science (SCS) efficaciously applied KM techniques to decorate collaboration, innovation, and facts sharing. Through a multi-pronged approach, SCS leveraged technology, culture, and procedures to create an ecosystem that encourages understanding sharing, subsequently contributing to the educational and lookup excellence of the institution.

In this case study, we delve into the experience of Carnegie Mellon University's School of Computer Science in imposing KM strategies, and how these techniques have positively impacted the institution's operations. Carnegie Mellon University's School of Computer Science is famed for its leadership in computer science and information technology. The diverse range of programs and the highly collaborative research environment has created a wealth of knowledge that, when managed effectively, can lead to innovation and excellence. SCS recognized the need for a robust technological infrastructure to support KM initiatives. This included the implementation of a Knowledge Management System (KMS) that facilitated knowledge capture, storage, retrieval, and sharing. The KMS allowed faculty and staff to contribute their expertise and access the collective knowledge of the institution (Chen et al., 2018) [33].

Some of the benefits of the implementation of KM include the following. A culture of collaboration was cultivated through training, workshops, and incentives. Faculty and staff were encouraged to participate in knowledge sharing activities and rewarded for their contributions. This culture shift was pivotal in breaking down silos and fostering interdisciplinary collaborations (Nonaka & Takeuchi, 1995) [13]. SCS reviewed and reengineered present strategies to contain KM practices. Workflows have been optimized to make certain that information sharing used to be an indispensable phase of day-by-day operations. The organization additionally established processes for knowledge creation, capture, validation, and dissemination (Davenport & Prusak, 1998) [16].

The KMS, coupled with the sub-culture of collaboration, resulted in accelerated interdisciplinary lookup and expertise sharing. Faculty and workforce pronounced multiplied possibilities for joint tasks and a deeper feel of belonging to a collaborative community. Access to a repository of institutional know-how enabled researchers to construct upon current work, accelerating the tempo of innovation. This was evident in the increased number of patents and groundbreaking research publications. The KMS provided decision-makers with real-time data and information, aiding in strategic planning and resource allocation. Faculty and administrators reported a more data-driven decision-making process. The implementation of KM at SCS was not without challenges. Some faculty members were initially resistant to change and reluctant to share their expertise. It was important to address these concerns through training and by highlighting the benefits of knowledge sharing. Carnegie Mellon University's School of Computer Science's successful implementation of KM strategies has significantly enhanced collaboration, innovation, and decision-making processes. The multi-pronged approach, involving technology, culture, and process re-engineering, serves as a model for other educational institutions looking to leverage KM to achieve academic excellence.

3.4 Case Study 4: Imperial College London's Data Science Institute

This case examines how Imperial College London's Data Science Institute efficaciously applied KM strategies, focusing on technology, culture, and processes, to smooth flow of research, collaboration, and innovation inside the institute. The article sheds light on the

elements contributing to the profitable KM implementation and its effect on the educational community.

In the era of big data and data-driven decision-making, academic institutions like Imperial College London recognize the significance of KM in advancing research and scholarship. Imperial College London's Data Science Institute (DSI) serves as a prime example of how a world-class research institute can effectively employ KM strategies to promote collaboration, innovation, and knowledge sharing.

The Data Science Institute at Imperial College London is renowned for its leadership in data science, artificial intelligence, and machine learning. The institute hosts numerous varieties of programs and multidisciplinary collaborations, which generate a large extent of facts and knowledge. The DSI identified the significance of a superior technological infrastructure to help KM initiatives. This included the development of a Knowledge Management System (KMS) that efficiently captured, stored, and facilitated the retrieval and sharing of research insights and findings. The KMS served as a centralized repository for research data, publications, and expertise (Alavi & Leidner, 2001) [28].

The DSI focused on fostering a culture of collaboration through a combination of incentives, training, and a shared mission. Faculty, researchers, and personnel had been inspired to actively have interaction in know-how sharing activities. Cross-disciplinary events and collaborative projects were promoted, breaking down silos and fostering a sense of shared commitment to research excellence (Nonaka & Takeuchi, 1995) [13]. DSI undertook a comprehensive review of existing processes to incorporate KM practices. Workflows were adapted to ensure that knowledge sharing became an integral part of daily operations. The institute also established processes for knowledge creation, validation, and dissemination, emphasizing the importance of effective data management and open science principles (Davenport & Prusak, 1998) [16]. Some of the benefits of KM in the college include the following. The KMS, coupled with a collaborative culture, resulted in an increase in multi-disciplinary research and cross-institute collaboration (Velazco et al., 2021). Researchers reported more opportunities for shared projects and an accelerated pace of innovation in data science and related fields.

The KMS provided researchers and faculty members with easy access to previous work, facilitating the building of new research upon existing findings. This led to an increase in research productivity, more publications, and contributions to cutting-edge research. The KMS allowed administrators and faculty members to access real-time data and knowledge, enabling data-driven decision-making processes related to research directions and resource allocation. The successful implementation of KM at DSI was not without challenges. Resistance to change and initial reluctance to share knowledge were common issues. These were addressed through continuous training and awareness campaigns that emphasized the benefits of knowledge sharing. Imperial College London's Data Science Institute's effective implementation of KM strategies has significantly enhanced collaboration, innovation, and decision-making processes. The combination of technology,

cultural shift, and process re-engineering serves as an exemplary model for other academic institutions aiming to leverage KM to advance research and academic excellence.

Table 1. Key Lessons Learned from the Four Case Studies on Knowledge Management Implementation

Name and Type of Organizations	Focus and emphasis			
Xerox Corporation (International enterprise)	Leadership commitment.	Aligned with the company's strategic objectives.	Collaborative culture.	Continuous learning and improvement.
Siemens AG (Multinational conglomerate)	Leadership involvement, Clear goals and objectives.	Aligned with strategic priorities.	Knowledge sharing culture.	Continuous learning and improvement
Carnegie Mellon University's School of Computer Science (SCS) (Educational school)	Reviewed and re-engineered strategies.	Focused on technological infrastructure.	Culture of collaboration.	Established KM processes and system.
Imperial College London's Data Science Institute (Academic institution)	Recognized the significance of KM, emphasized innovation.	Emphasized advance technological infrastructure.	Culture of collaboration, shared commitment.	Continuous training and campaigns, established procedures, and workflow.

Overall, Xerox, Siemens, Carnegie Mellon University's School of Computer Science (SCS) and Imperial College London's Data Science Institute had similar goals and emphasis but took different approaches to implementing their KM programs. Those companies recognized the importance of KM implementation, provided leadership, established technological infrastructure, aligned their strategies, taught on collaborative culture, provided continuous training, and faced challenges in adoption and integration (Table 1). Other lessons from the case studies on KM strategies are the importance of creating a culture of knowledge-sharing, using technology to enable knowledge-sharing, measuring the impact of KM, identifying, and prioritizing knowledge, using a centralized KM system, and investing in training and development programs.

4. CONCLUSION

In conclusion, "The Evolving Trend, Development, and Strategies for Effective

Knowledge Management in the Digital Era" provides a comprehensive overview of the dynamic landscape of KM in the modern digital age. Through an in-depth analysis of the latest trends, developments, and strategies, the article sheds light on the challenges and opportunities associated with effective KM in today's rapidly changing technological environment. KM has become increasingly important for organizations in many aspects and the digital era has resulted in changes in the way knowledge is created, managed, accessed, and shared. The advancement of KM in the digital era is driven by the need to manage vast data and the need to be more responsive to the evolving market environment. To ensure efficient KM in the digital era, some of the KM strategies for best practice need to consider the highlights for organizations to develop a culture of knowledge sharing, invest in KM tools for effective implementation. One key takeaway from this article is the need for organizations to continuously adapt and evolve their KM practices keeping pace with the ever-evolving digital landscape. The article discusses the importance of leveraging on advancing technologies, such as artificial intelligence and big data analytics, to effectively capture, organize, and utilize knowledge within organizations. Furthermore, the proposed outline of the paper with a focus on KM strategies will contribute towards organizational implementation practices. In terms of implication, an area for consideration for future research is KM and artificial intelligence. There is a need to investigate how organizations can use artificial intelligence to manage knowledge more effectively. With the right approach, organizations can benefit and gain a competitive advantage in managing their knowledge. Implementing KM practices is a lesson smart organization are discovering and learning again. As we move forward, the article suggests that organizations need to be agile, adaptive, and forward-thinking in their KM strategies, to stay competitive in the fast-paced digital era. It calls for a proactive and strategic approach to managing knowledge, leveraging technology, and nurturing a knowledge-sharing culture. In conclusion, “Knowledge Management in the 21 st-Century: Trends, Developments, and Strategies" underscores the importance of embracing digital transformation, adopting innovative technologies, and nurturing a culture of knowledge-sharing, to effectively manage knowledge in todays in state-of-the-art dynamic commercial enterprise environment. It serves as a valuable resource for organizations and practitioners seeking to navigate the complexities of KM in the digital era and stay ahead in the swiftly altering landscape of information and technology.

REFERENCES

[1] Idrees, H., Xu, J., Haider, S. A., & Tehseen, S. (2023). A systematic review of knowledge management and new product development projects: Trends, issues, and challenges. *Journal of Innovation & Knowledge,* 8(2), 100350.

[2] Al-Shahrani, M. M. (2019). Trends in knowledge management processes and practices. *Journal for Research on Business and Social Science.* ISSN (Online) 2209-7880, 2(12).

[3] Usman, M., Naveed, R. T., Iqbal, A., Mustafa, G., & Anwar, A. (2020). The Importance and Implication of Knowledge Management and Its Impact on Organizational Performance. *Abasyn University Journal of Social Sciences*, *13*(1).

[4] Manesh, M. F., Pellegrini, M. M., Marzi, G., & Dabic, M. (2020). Knowledge management in the fourth industrial revolution: Mapping the literature and scoping future avenues. *IEEE Transactions on Engineering Management, 68*(1), 289-300.

[5] Girard, J. & Girard, J. (2015). Defining knowledge management: Toward an applied compendium. *Journal of Applied Knowledge Management*. Vol 3, Issue 1.

[6] Sharma, V. (2014). *Knowledge Management practices at Microsoft.* https://www. slideshare.net/ VanishreeSharma/km-practices-at

[7] Alshammari, A., Elamer, A., & Brooks, C. (2020). *Knowledge Management and Firm Performance: A Meta-analysis.* https://www.sciencedirect.com/ science/article/pii/ S030 504 8320 303 322

[8] Toma, S. G. (2006). Defining management for the twenty-first century. *The AMFITEATRU ECONOMIC Journal*, *8*(19), 122-125.

[9] Drucker, P. F. (1989), *The New Realities*, Harper-Collins Publishers, New York, US.

[10] Roshchin, I., Pikus, R., Zozulia, N., Marhasova, V., Kaplinskiy, V., & Volkova, N. (2022). Knowledge management trends in the digital economy age. *Postmodern Openings, 13*(3), 346-357.

[11] Ganapathy, S., Mansor, Z., & Ahmad, K. (2020). Trends and challenges of knowledge management technology from Malaysia's perspective. *International Journal on Advanced Science Engineering Information Technology.* Vol. 10. No. 4. ISSN: 2088-5334.

[12] Mohajan, H. (2017). The roles of knowledge management for the development of organizations. *Journal of Scientific Achievements.* Vol. 2, No. 2, p. 1 p 27.

[13] Nonaka, I., & Takeuchi, H. (1995). *The knowledge-creating company: How Japanese companies create the dynamics of innovation.* Oxford University Press.

[14] Kinyata, G. L. (2014). The role of knowledge management in higher education institutions: A case study from Tanzania. *International Journal of Management, Knowledge and Learning,* 3(1), 43-58.

[15] Igbinovia, M. O., & Ikenwe, I. J. (2017). Knowledge management: processes and systems. Information Impact: *Journal of Information and Knowledge Management*, 8(3), 26-38.

[16] Davenport, T. & Prusak, L. (1998). *Working knowledge: How organizations manage what they know.* 10.1145/348772.348775. Harvard Business School Press. Retrieved from https://www. Research gate.net/publication/229099904.

[17] Sardjono, W., & Firdaus, F. (2020). Readiness model of knowledge management systems implementation at the higher education. *ICIC Express Letters, 14*(5), 477-487.

[18] Bellinger, G., Castro, D., & Mills, A. (2004). *Data, information, knowledge, and wisdom.* http:// outsights.com/ systems/dikw/dikw.htm

[19] Baptista Nunes, J. M., Kanwal, S., & Arif, M. (2017). *Knowledge management practices in higher education institutions: A systematic literature review*. library.ifla.org. http://creativecommons. org/ licenses/by/4.0

[20] Ismail, H. & Abdullah. R. (2016). Knowledge management best practice in higher learning institution: a systematic literature review. *Journal of Theoretical and Applied Information Technology*. Vol.90. No.2. ISSN: 1992-8645

[21] Bughin, J., Catlin, T., Hirt, M., & Willmott, P. (2018). McKinsey & Company. Why digital strategies fail. *McKinsey Quarterly.*

[22] Dalkir, K., & American Psychological Association. (2011). *Knowledge management in theory and practice* (2nd Ed). ISBN-13: 978-0-262-31058-1. MIT Press.

[23] Alavi, M. & Leidner, D.E. (2001). Knowledge management and knowledge management systems: conceptual foundations and research issues. *MIS Quarterly*, *25*, 107-136. http://dx.doi.org/ 10. 2307/ 32 50961

[24] Wasko, M. M., & Faraj, S. (2005). Why should I share? Examining social capital and knowledge contribution in electronic networks of practice. *MIS quarterly*, 35-57.

[25] Alvesson, M., & Karreman, D. (2011). Varieties of discourse: On the study of organizations through discourse analysis. *Human Relations*, *64*(11), 1305-1339.

[26] Chen, C. (2016). Knowledge management practices in supporting innovation: An integrative framework. *Information & Management*, 53(6), 643-657

[27] Hlatshwayo, M. (2019). Information and knowledge management. *Journal of Information & Knowledge Management,* 1-12.

[28] Powers, V. J. (1999). Xerox creates a knowledge-sharing culture through grassroots efforts. *Knowledge Management in Practice*, 18(1), 1-4.

[29] Hickins, M. (2013). *Xerox shares its knowledge*. The knowledge management yearbook 2000-2001 (pp. 98-107). Routledge.

[30] Sveiby, K. E. (2001). A knowledge-based theory of the firm to guide in strategy formulation. *Journal of Intellectual Capital*, 2(4), 344-358

[31] Alavi, M., & Tiwana, A. (2002). Knowledge integration in virtual teams: The potential role of KMS. *Journal of the American Society for Information Science and Technology*, *53*(12), 1029-1037.

[32] Riemer, K. (2001). Knowledge management at Siemens: Making sense of the organization. *International Journal of Technology Management*, 22(3/4), 316-331.

[33] Chen, D., Preston, D. S., & Xia, W. (2018). Toward an Intelligent Campus: A Knowledge Management Framework. In *Proceedings of the 51st Hawaii International Conference on System Sciences.*

Journal of Science and Development Economics
Nam Can Tho University

Website: jsde.nctu.edu.vn

Renewable energy generation and energy efficiency in seaports: a focus on the Malaysian maritime industry

Thiagarajan Marappan[1*], M. Vikneswary Suresh[1]

[1]*Malaysia University of Science and Technology, Malaysia*

[*]*Corresponding author: Thiagarajan Marappan (email: thiagarajan@must.edu.my)*

Received: November 10, 2023
Revised: November 30, 2023
Accepted: December 5, 2023

Keywords: *green port, maritime transportation, renewable energy, sustainable energy, technological solutions*

Từ khóa: *bền vững, cảng xanh, giải pháp công nghệ, năng lượng tái tạo, vận tải biển*

ABSTRACT

The global maritime industry plays a crucial role in international trade and transportation, with seaports serving as vital hub. In recent years, there has been growing interest in transitioning seaports towards sustainable practices by integrating renewable energy sources for power generation and consumption. Additionally, the global community is witnessing a shift towards a fourth industrial revolution, which has sparked a new wave of energy generation and consumption transformation. Peak shaving can balance load demand and facilitate the participation of small power units in generation based on renewable energy sources. In this regard, many approaches have been introduced, such as solar Power Photovoltaic (PV), wind energy, biofuel, biomass, and Battery Energy Storage (BES) which act as Energy Storage System (ESS). To assess the viability of renewable energy application in ports, various technological solutions and innovative practices will be analysed. Furthermore, this article presents a focused analysis of the Malaysian context, which encompasses a thriving maritime sector, to offer valuable perspectives on the utilisation of renewable energy solutions by seaports in this geographical area. This article seeks to make a scholarly contribution to the ongoing discussion on sustainable port operations within the broader context of the global shift towards renewable energy sources, by conducting a comprehensive examination of technological advancements and sustainable practices.

TÓM TẮT

Ngành hàng hải toàn cầu đóng một vai trò quan trọng trong thương mại và vận tải quốc tế, trong đó các cảng biển đóng vai trò là trung tâm quan trọng. Trong những năm gần đây, mối quan tâm ngày càng tăng đối với việc chuyển đổi cảng biển theo hướng bền vững bằng cách tích hợp các nguồn năng lượng tái tạo để sản xuất và tiêu thụ điện. Ngoài ra, cộng đồng toàn cầu đang chứng kiến sự chuyển dịch sang cuộc cách mạng công nghiệp lần thứ tư, điều này đã tạo ra một làn sóng mới về chuyển đổi sản xuất và tiêu thụ năng lượng. Việc cạo đỉnh có thể cân bằng nhu cầu phụ tải và tạo điều kiện thuận lợi cho các đơn vị điện lực nhỏ tham gia phát điện dựa trên các nguồn năng lượng tái tạo. Về vấn đề này, nhiều phương pháp đã được đưa ra, như Quang điện mặt trời (PV), năng lượng gió, nhiên liệu sinh học, sinh khối và Bộ lưu trữ năng lượng pin (BES) hoạt động như Hệ thống lưu trữ năng lượng (ESS). Để đánh giá tính khả thi của việc ứng dụng năng lượng tái tạo tại các cảng, nhiều giải pháp công nghệ và thực tiễn đổi mới sẽ được phân tích. Hơn nữa, bài viết này trình bày phân tích tập trung về bối cảnh của Malaysia, trong đó bao gồm lĩnh vực hàng hải đang phát triển mạnh, nhằm đưa ra những quan điểm có giá trị về việc sử dụng các giải pháp năng lượng tái tạo của các cảng biển trong khu vực địa lý này. Bài viết này tìm cách đóng góp về mặt học thuật cho cuộc thảo luận đang diễn ra về hoạt động cảng bền vững trong bối cảnh rộng lớn hơn của sự thay đổi toàn cầu hướng tới các nguồn năng lượng tái tạo, bằng cách tiến hành kiểm tra toàn diện các tiến bộ công nghệ và thực tiễn bền vững.

1. INTRODUCTION

Seaports play a pivotal role in global trade and serve as interfaces between land and sea transportation, facilitating the transfer of cargo from ships to trucks, trains, and vice versa (Gurzhiy et al., 2021) [1]. Seaports generate employment opportunities directly and indirectly, enhance the competitiveness of domestic industries, and foster economic growth by attracting foreign investment and trade (Caliskan, 2022) [2]. Seaports have a strategic importance for national security. They are crucial for the movement of military supplies and personnel during times of conflict or humanitarian crises (Anser et al., 2020) [3]. However, the operations of seaports also present significant environmental challenges posit that seaport activities, particularly those involving ships and trucks, release pollutants into the air, including sulphur dioxide, nitrogen oxides, and particulate matter. These emissions can lead to poor air quality and have adverse

health effects on nearby communities. Seaports can discharge pollutants, such as ballast water, oil, and chemicals, into nearby water bodies, causing harm to aquatic ecosystems and marine life (Melnyk et al., 2023) [4]. Inadequate wastewater treatment facilities can exacerbate water pollution (Samsudin et al., 2016) [5]. According to Yadav et al., (2021) [6], addressing these environmental challenges associated with seaport operations requires a multi-faceted approach, including the adoption of cleaner technologies, improved waste management, sustainable urban planning, and regulatory measures to reduce emissions and pollution. In addition, there is a growing emphasis on green port initiatives and the development of eco-friendly infrastructure to mitigate the environmental impact of ports while continuing to support global trade.

2. RENEWABLE ENERGY AND PORTS

Amid the current global climate crisis and the urgent need for energy efficiency, there has been an increase in concerns regarding the environmental impacts of port development and operations. The environmental impact of traditional seaport operations, characterized by emissions from vessel, handling equipment, and energy-intensive facilities, has raised concerns about sustainability. The renewable energy generation and consumption in seaports have gained prominence as an innovative and sustainable solution to address the dual challenges of environmental stewardship and energy security. This transition not only aligns with global effort to combat climate change, such as the Paris Agreement, but also offers numerous perceived benefits. These advantages include reduced carbon emissions, improved air quality, lower operating costs, and increased energy resilience.

The International Maritime Organisation (IMO) is a United Nations specialised agency tasked with the global regulation of shipping. Amid the urgent matter of climate change and its correlation with maritime operations, the IMO has established ambitious objectives to mitigate GHG emissions. IMO has set a goal of achieving a 50% reduction in annual total GHG emissions and a 70% reduction in carbon dioxide (CO2) emissions from transport operations by 2050. Figure 1 illustrate the IMO Strategy on Reduction of GHG Emissions from ships.

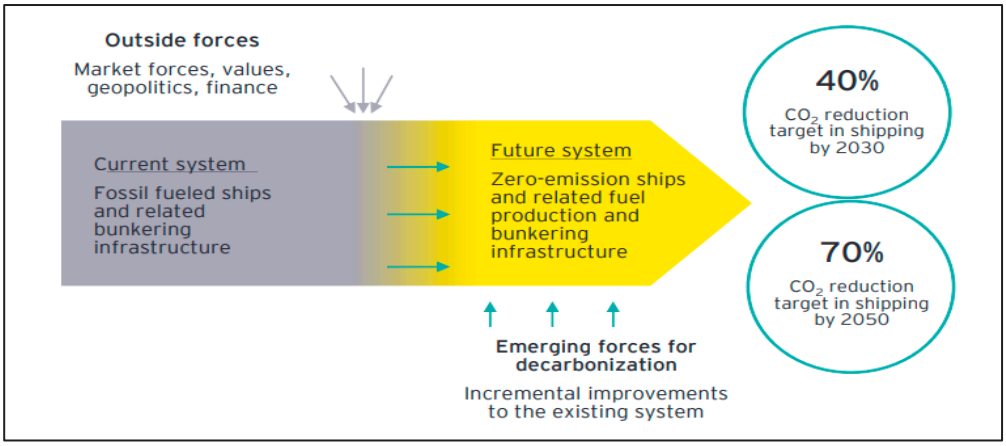

Figure 1. Imo strategy on reduction of ghg emissions from ships

Source: World Economic Forum, 2023

In addition, the European Sea Ports Organisation (ESPO) Report 2022 presents the annual environmental benchmarks for the European port sector, based on data from ports that are members of the EcoPorts Network. The results of an analysis of several environmental performance indicators, which are presented in Figure 2, lead to the conclusion that the most pressing environmental concerns in the year 2019 are related to climate change, air quality, and energy consumption. Therefore, reducing the negative effects that port operations have on the surrounding environment should be the primary motivation for adopting renewable energy sources in ports.

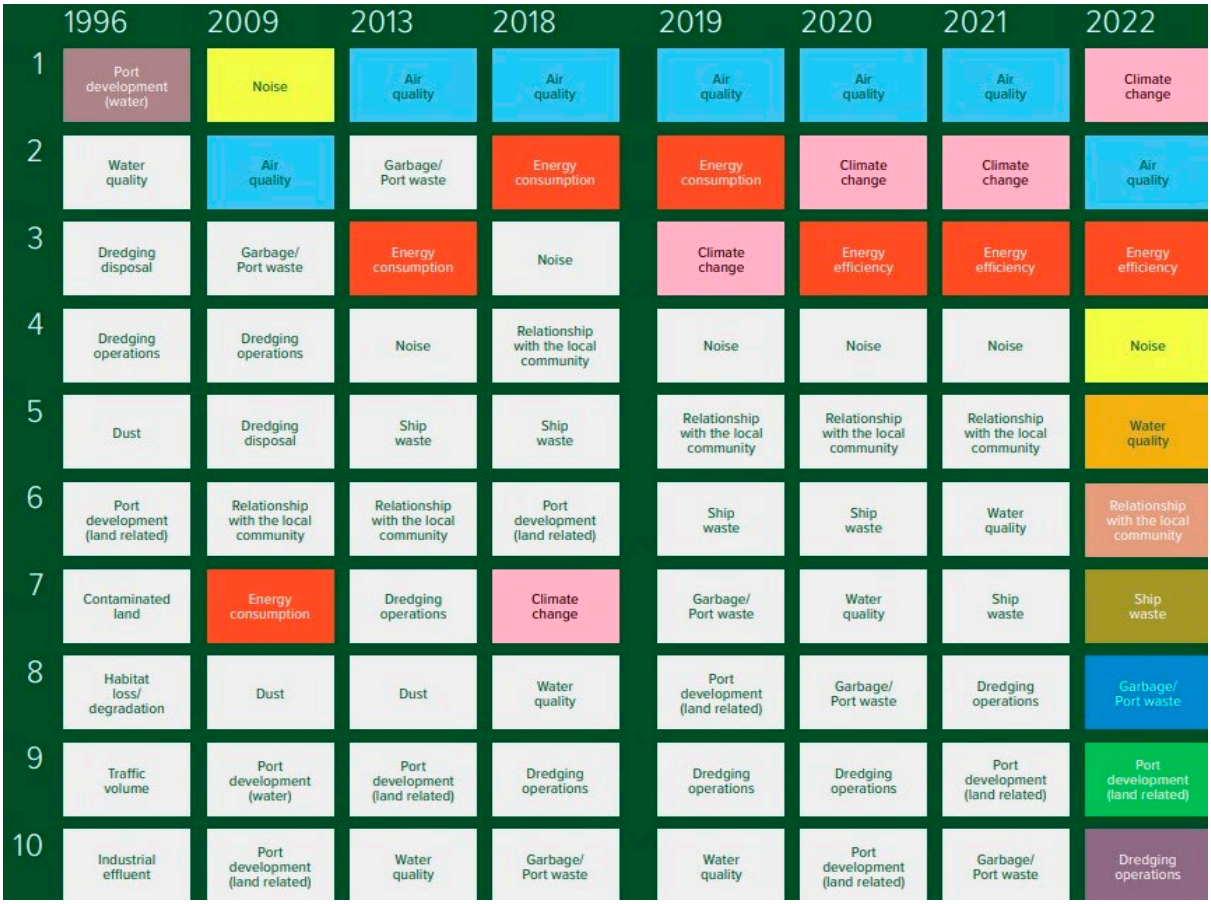

Figure 2. Top 10 Environmental Priorities of the Port Sector over the Year

Source: ESPO

3. RENEWABLE ENERGY TECHNOLOGIES IN PORTS

Renewable energy technologies in seaports have gained significant attention and implementation in recent years. To ensure sustainability, a port must effectively monitor and reconcile three fundamental aspects: environmental quality, economic prosperity, and societal well-being. According to Sarmento et al. (2018), the economic benefits of renewable energy, such as reduced energy costs over time, are a significant motivation for ports to transition

to cleaner energy source. This is also supported by Guo et al., (2019) that the use of renewable energy in ports can improve energy security by lowering reliance on fossil fuels and increasing the availability of diverse energy sources. The following are the technologies that offer various benefits, including reducing GHG emissions, lowering operational costs, enhancing energy security, and promoting sustainable development (Rajaram et al., 2019).

3.1 Power Photovoltaic

Power Photovoltaic (PV) has gained prominence in many ports as a clean and sustainable energy source. Ports with extensive rooftop areas, such as the Port of Los Angles, have installed solar panel to generate electricity for their operations. In addition, floating solar farms on water bodies within port premises, such as those seen in the Port of Singapore, are emerging as viable option for maximizing solar power generation. Solar energy is the most promising source of clean, renewable energy and has the greatest potential to solve the world's energy problems. An energy source that uses the radiant light emitted from the sun is called solar energy and can be converted into electrical energy using a device called photovoltaic cell (Ong., H.C et al., 2011). Photovoltaic cells convert sunlight directly into electricity without creating air or water pollution. Solar energy has been identified the cleanest renewable energy source.

3.2 Wind Energy

Wind turbines have found their way into seaport area, taking advantage of the often-windy coastal locations. The Port of Rotterdam, Europe's largest port, features a dedicated wind farm that supplies renewable energy to the seaport's infrastructure. These turbines not only provide clean energy but also serve as iconic landmarks in the port landscape.

3.3 Biomass and Biofuels

Several seaports are exploring the use of biomass and biofuels to replace traditional fossil fuels. The Port of Gothenburg in Sweden has invested in biogas-powered trucks and machinery for cargo handling. In addition, biofuel blends are becoming increasingly common for vessels docking at ports worldwide, reducing emissions from marine transportation. Biomass is the mass of combustible materials of organic origin from any source such as plants, bio wastes or process wastes (Acciaro et al., 2014) [7]. Biomass gives energy via different conversion technologies in the form of heat or electricity, or into other forms such as liquid biofuel or combustible biogas. The conversion technologies for solid biomass resources into heat, power and combine heat and power (CHP) can be classified into two general groups, i.e., combustion and gasification. Another method for conversion of solid biomass is anaerobic digestion. Biomass also provides transportation fuels such as bioethanol and biodiesel.

3.4 Battery Storage

Battery Energy Storage System (BESS) are gaining popularity in ports to store excess renewable energy for later use. These systems help ensure a stable and reliable power supply, particularly during peak demand periods or when renewable generation is low. The Port of Los Angeles, for instance, has deployed a significant BESS to enhance grid stability (Green Ports Gateways to Europe, 2020) [8].

3.5 Electric Vehicles (EVs)

Many ports are electrifying their vehicle fleets by introducing electric cargo handling

equipment and establishing EV charging infrastructure. This reduces emissions from port activities and contributes to the electrification of transportation within the port vicinity (National Energy Transition Roadmap, 2023) [9]

Seaports are major sources of GHG emissions because of the heavy use of diesel-powered equipment, such as cranes, trucks, and ships. Thus, transitioning to renewable energy sources, such as wind, solar, and hydroelectric power, can significantly reduce carbon emissions associated with seaport operations. However, alongside the promising benefits, there are also perceived risks and challenges associated with the integration of renewable energy into seaport operations. These include initial capital investments, grid integration complexities, variability in energy production, and potential disruptions to established supply chains. Therefore, a comprehensive understanding of dynamics, opportunities, and potential risk of renewable energy adoption in seaports is crucial for stakeholders in the maritime industry, policymakers, and researchers (Yang et al., 2021) [10].

4. RENEWABLE ENERGY ADOPTIONS SEAPIRTS

The need for renewable energy solutions in seaports is driven by the imperative to reduce carbon emissions and enhance sustainability in the maritime industry. However, the high level of activity in seaports also leads to significant environmental impacts, primarily due to the reliance on fossil fuels. Renewable energy solutions offer a way to mitigate these impacts. Therefore, using renewable energy for electricity generation and electrifying seaport equipment can reduce local air pollution by eliminating diesel emissions.

Seaports are required to meet global and regional regulations, such as the International Maritime Organization's (IMO) sulphur cap and emissions reduction targets, require seaports and ships to reduce their environmental footprint (Ariffin et al., 2022) [11]. Furthermore, renewable energy adoption in seaports can position them as leaders in sustainable maritime operations, attracting environmentally conscious shipping companies and shippers while bolstering their reputations as responsible stewards of their local ecosystems.

Notable case studies such as port of Los Angeles and port of Rotterdam demonstrate the feasibility of these initiatives as well as the benefits that can be gained from participating in them. Even though there are still obstacles to overcome, the overall outlook for renewable energy in seaports is optimistic, which is in line with efforts being made all over the world to combat climate change. This overview provides insights into the state of renewable energy adoption in seaports globally, drawing upon recent developments and notable examples.

Table 1. Current Ongoing Renewable Energy Projects in Ports

Country	Port name	Technologies
United Arab Emirates (UAE)	Port of Jebel Ali	Solar Photovoltaic
Sweden	Port of Gothenburg	Hydrogen production facility

Country	Port name	Technologies
Australia	Por of Brisbane	The Solar Energy Initiative
The Netherlands	Port of Rotterdam	Wind farm & Biofuel refinery
Finland	Port of Helsinki	Wind Power
Germany	Port of Hamburg	Solar Photovoltaic
New Zealand	Ports of Auckland	Solar Photovoltaic
China	Port of Qingdao	Hydrogen
Italy	Port of Genoa	Solar, biomass, wind, geothermal energy
Fiji	Ports of Fiji	Solar Photovoltaic
Sri Lanka	Port of Colombo	Solar Photovoltaic

Source: Author

5. RENEWABLE ENERGY AND EVIDENCE FROM MALAYSIAN SEAPORT

Malaysia is one of the world's fastest-growing, competitive economies with a significant position in the Southeast Asian region attributed to its location at the confluence of the intercontinental and intra-Asian maritime trade routes going through the Strait of Malacca. Hence, it is no surprise that Malaysia houses one of the biggest port facilities in the world. It is also a transshipment hub of the Asian region and a preferred point of entry into the Southeast Asian Market. Boasting a highly developed maritime shipping sector, Malaysia has been ranked by UNCTAD as the fifth-best linked country in the world, in terms of shipping trade route connectivity, better than the developed economies of Germany and the Netherlands. There are six major ports in Malaysia that are currently being identified as the major trade route for international trade namely, Port Klang, Port of Tanjung Pelepas, Port of Johor, Port of Bintulu, and Port of Kuantan.

Malaysia is located between 1 degree and -7 degree in North latitude and between 100 degree and 120 degree in East longitude, which is second largest solar radiation region (Samsudin et al., 2016) [5]. Therefore, there is a large potential for photovoltaic energy to be absorbed by photovoltaic cells in Malaysia. The average daily solar radiation in Malaysia is within the range of 4.12 – 5.56 kWh/m^2. The highest solar radiation was estimated to be 6.8 kWh/m^2/day in August and November, whereas the lowest was found to be 0.61 kWh/m^2/day in December (Mansor, 2014). According to the International Energy Agency, it is estimated that approximately 141 GWh of electricity was produced in 2013, which 0.1% of the total electricity generation in Malaysia. The installed capacity of solar energy in Malaysia is 0.07 GW (Mekhilef et al., 2014).

In terms of, the wind speed is generally low over Malaysia and there is no remarkable electricity has been generated using wind energy in Malaysia (Bose et al. 2019). However, some areas of Malaysia have experienced a considerable wind speed for wind power generation.

Table 2. Wind speed and power density in few prospective areas of Malaysia

Area	Power density (W/m^2) at 10 m height	Power density (W/m^2) at 65 m height	Probity of speed above 2.5 m/s (%)	Turbine operating hours
Kota Bharu	11.058	45.015	38.79	3398
Kuala Terengganu	7.367	30	28.37	2485
Langkawi Island	5.822	23.7	20.75	1818
Mersing	17.013	69.257	58.04	5084
Miri	7.197	29.298	28.85	2527

Source: Samsudin et al., (2016) [5]

Power generation in Malaysia heavily relies on three fossil fuel sources, namely coal, natural gas, and fuel-oil. However, the current power generation pattern is not sustainable due to the adverse environmental impacts and the depletion of fossil fuel reserves. Fortunately, Malaysia possesses abundant renewable energy resources, particularly biomass, solar, small hydro, and ocean energy. Considering this, the utilization of renewable energy and the implementation of energy efficiency measures emerge as crucial tools in achieving sustainable energy solutions.

The reserves of the main fossil fuel sources in Malaysia are estimated at 1.94 billion tons for coal, 0.64 billion tonnes for crude oil, and 2784 billion cubic meters for natural gas. The reserves of fossil fuels will be completely exhausted in the near future, e.g. natural gas by 40 years and oil by 29 years from now. Although the country has a relatively big amount of coal reserves, they are concentrated in Sarawak and Sabah which have inadequate infrastructure and high extraction cost. Malaysia endowed with huge hydropower resources that can generate as much as 29,000 MW of electricity. Malaysia produces large amount of palm oil biomass annually, which can be the major contributor of renewable energy. The wind speed is generally lower over Malaysia and there is no remarkable potential of wind energy (Samsudin et al., 2016) [5].

While the adoption of renewable energy in ports is growing, challenges remain. High upfront costs, regulatory hurdles, and the intermittency of renewable sources are among the key obstacles. However, as technology advances and economies of scale are realized, these challenges are likely to diminish. The global trend toward decarbonization and the imperative to combat climate change ensure that renewable energy adoption in ports will continue to expand (Samsudin et al., 2016) [5].

Malaysia Government Initiatives:

The 12th Malaysia Plan, spanning the years 2021 to 2025, and the National Energy Policy 2022-2040, have laid the groundwork while the NETR will ensure Malaysia forges ahead in this

transformative journey. Against the backdrop of a dynamic global energy landscape mired with the energy trilemma of security, affordability, and sustainability at its core, the world is racing for solutions. Malaysia too is resolute in overcoming these challenges and the NETR demonstrates our unwavering determination in this regard. Reducing Malaysia's carbon footprint is one of the catalysts to transforming the economy on a more sustainable path. It is also an agenda to generate new sources of growth, creating business and trade opportunities, and consequently, knowledge workers (National Energy Transition Roadmap, 2023) [9]. Malaysia's transition to a low-carbon, climate resilient economy is driving industry leaders to prioritize focus on reducing greenhouse gas (GHG) emissions and investing in low emissions technologies. In supporting the global decarbonization initiatives, Malaysia has set a reduction target of up to 45% of GHG emission intensity to GDP by 2030 and this demands the participation of economic stakeholders across all sectors (National Energy Policy, 2021) [12]. Malaysian ports can benefit from collaborating with international ports that have successfully adopted renewable energy solutions. Participation in international initiatives and organization can provide access to funding and expertise. The Malaysian government can further promote renewable energy adoption in ports by offering incentives, subsidies, and clear regulations. Encouraging partnership between port authorities and private companies can facilitate investment in renewable energy projects. Ten flagship catalyst projects of the NETR, which cover six energy transition levers namely, energy efficiency (EE), renewable energy (RE), hydrogen, bioenergy, green mobility, and carbon capture, utilisation, and storage (CCUS) was launched in July, 2023. These flagship projects are expected to attract investment of more than RM25 billion, create 23,000 job opportunities and reduce GHG emissions by more than 10,000 Gg CO2eq per year (National Energy Transition Roadmap, 2023) [9].

6. CONCLUSION

The adoption of renewable energy sources in ports is a multifaceted process driven by environmental concerns, cost savings, and energy security considerations. While there are challenges related to initial costs, intermittency, and grid integration, ongoing advancements in technology and supportive policy frameworks are encouraging ports to make the transition. As research in this field continues to evolve, it is essential to explore innovative solutions and share best practices to further promote the sustainability and environmental performance of ports. Over the next three decades, Responsible Transition pathway sets the direction to meet growing energy needs and reduce GHG emissions. Malaysia will focus on improving energy efficiency, enhancing RE and bioenergy, reducing GHG emissions, greening mobility, accelerating innovation to commercialise hydrogen and CCUS technologies as well as strengthening energy infrastructure. These actions will be accompanied by strategies to unlock capital flows in support of the energy transition with energy security as the cornerstone. In terms of prospects, it is expected that the use of RETs in green ports will continue to increase in the coming years. This trend will be driven by the growing awareness of the need for sustainable

development, decreasing the installation cost of RETs, and the availability of supportive policies and incentives. The development of innovative technologies such as floating solar panels, wind-powered shipping, and green hydrogen production will also open new opportunities for the application of renewable energy in green ports. Additionally, the integration of RETs with smart port technologies and digital solutions such as artificial intelligence, blockchain, and the Internet of Things will enable better monitoring, management, and optimization of energy consumption in green ports.

7. ACKNOWLEDGEMENTS

The author expresses his gratitude to the Malaysia University of Science and Technology for their support in the publication of this conference paper.

REFERENCES

[1] Gurzhiy, A., Kalyazina, S., Maydanova, S., & Marchenko, R. (2021). Port and City Integration: Transportation Aspect. *Transportation Research Procedia, 54*, 890–899. https://doi.org/10.1016/j.trpro.2021.02.144

[2] Caliskan, A. (2022). Seaports participation in enhancing the sustainable development goals. *Journal of Cleaner Production, 379*, 134715. https://doi.org/10.1016/J.JCLEPRO.2022.134715

[3] Anser, M. K., Yousaf, Z., & Zaman, K. (2020). Green Technology Acceptance Model and Green Logistics Operations: "To See Which Way the Wind Is Blowing." *In Frontiers in Sustainability* (Vol. 1). Frontiers Media S.A. https://doi.org/10.3389/frsus.2020.00003

[4] Melnyk, O., Onyshchenko, S., & Onishchenko, O. (2023). Development Measures to Enhance the Ecological Safety of Ships and Reduce Operational Pollution to the Environment. *Scientific Journal of Silesian University of Technology. Series Transport, 118*, 195–206. https://doi.org/10.20858/sjsutst.2023.118.13

[5] Samsudin, M. S. N., Rahman, M. M., & Wahid, M. A. (2016). Akademia Baru Power Generation Sources in Malaysia: Status and Prospects for Sustainable Development. *Journal of Advanced Review on Scientific Research ISSN* (Vol. 25, Issue 1).

[6] Yadav, H., Soni, U., & Kumar, G. (2021). Analysing Challenges to Smart Waste Management for a Sustainable Circular Economy in Developing countries: A fuzzy DEMATEL study. *Smart and Sustainable Built Environment, 12*(2), 361–384. https://doi.org/10.21203/rs.3.rs-263855/v1

[7] Acciaro, M., Ghiara, H., & Cusano, M. I. (2014). Energy management in seaports: A new role for port authorities. *Energy Policy, 71*, 4–12. https://doi.org/10.1016/j.enpol.2014.04.013

[8] Green ports gateways to Europe. (2020).

[9] National Energy Transition Roadmap. (2023).

[10] Yang, L., Danwana, S. B., & Yassaanah, F. L. I. (2021). An empirical study of renewable energy technology acceptance in ghana using an extended technology acceptance model. *Sustainability*

(Switzerland), 13(19). https://doi.org/10.3390/su131910791

[11] Ariffin, Z. Z., Isa, N., Lokman, M. Q., Ludin, N. A., Jusoh, S., & Ibrahim, M. A. (2022). Consumer Acceptance of Renewable Energy in Peninsular Malaysia. *Sustainability (Switzerland), 14*(21). https://doi.org/10.3390/su142114627

[12] National Energy Policy (2021).

Journal of Science and Development Economics
Nam Can Tho University

Website: jsde.nctu.edu.vn

Enhancing knowledge retention in it education: an investigation into the impact of improved microlearning course structures and segmentation strategies

Ang Ling Weay[1*], Sellappan Palanniappan[1]

[1]*School of Information Technology, Malaysia University of Science & Technology, Malaysia*

[*]*Corresponding author: Ang Ling Weay (email: dr.ang@must.edu.my)*

Received: November 10, 2023
Revised: November 30, 2023
Accepted: December 5, 2023

Keywords: *control group design, knowledge retention, learning performance, microlearning, pretest posttest*

Từ khóa: *hiệu suất học tập, kiến thức, kiểm soát, thiết kế, thử nghiệm*

ABSTRACT

This study introduced an improved structure for microlearning courses and investigated how the segmentation method within the microlearning system impacts students' ability to retain knowledge. We conducted an experiment before and after the treatment, involving 90 first-year students who were enrolled in an IT course, specifically the Information System Module. The students interacted with course content that was created using Gagné's Nine Events of Instruction. Both before and after the treatment, they were required to answer a set of 3 short-answer questions and 2 essay questions. Utilizing quantitative research methods, we assessed the effectiveness of the proposed microlearning segmentation process throughout an entire semester. We compared the final exam grades (post-test) with those of the mid-term exams (pre-test). The results indicate that students demonstrate the ability for self-directed learning in a self-regulated learning environment when they use the suggested segmentation approach. Furthermore, the findings show that students who followed the course design incorporating this segmentation process exhibited better knowledge retention compared to a traditional learning group. This increase in retention can be attributed to the key factors of intrinsic motivation, specifically perceived choice and perceived value.

TÓM TẮT

Nghiên cứu này đã giới thiệu một cấu trúc cải tiến cho các khóa học microlearning và điều tra xem phương pháp phân đoạn trong hệ thống microlearning tác động như thế nào đến khả năng

ghi nhớ kiến thức của sinh viên. Chúng tôi đã tiến hành một thử nghiệm trước và sau khi điều trị, với 90 sinh viên năm thứ nhất đăng ký khóa học công nghệ thông tin, đặc biệt là Mô-đun hệ thống thông tin. Học sinh tương tác với nội dung khóa học được tạo bằng 9 sự kiện giảng dạy của Gagné. Cả trước và sau khi xử lý, họ được yêu cầu trả lời một bộ 3 câu hỏi trả lời ngắn và 2 câu hỏi tiểu luận. Bằng cách sử dụng các phương pháp nghiên cứu định lượng, chúng tôi đã đánh giá tính hiệu quả của quy trình phân đoạn microlearning được đề xuất trong suốt toàn bộ học kỳ. Chúng tôi so sánh điểm thi cuối kỳ (kiểm tra sau) với điểm kiểm tra giữa kỳ (kiểm tra trước). Kết quả chỉ ra rằng sinh viên thể hiện khả năng tự học trong môi trường học tập tự điều chỉnh khi họ sử dụng phương pháp phân khúc được đề xuất. Hơn nữa, các phát hiện cho thấy rằng những sinh viên tuân theo thiết kế khóa học kết hợp quy trình phân khúc này cho thấy khả năng ghi nhớ kiến thức tốt hơn so với nhóm học tập truyền thống. Sự gia tăng khả năng ghi nhớ này có thể là do các yếu tố chính của động lực nội tại, cụ thể là sự lựa chọn được nhận thức và giá trị được nhận thức.

1. INTRODUCTION

The evolution and advancement of industries, coupled with shifts in labor market demands, have undergone significant transformations due to the rapid expansion of technology and society. In the era of Industry 4.0 (IR4.0), individuals must equip themselves with essential knowledge and skills while fostering a lifelong learning mindset [1],[2]. Both students and educators continuously update and enrich their knowledge systems, adapting to emerging technologies and evolving methods of applying acquired knowledge. Online teaching platforms encourage students to become self-reliant learners, demanding greater self-awareness. However, the drawback is that educators are unable to immediately gauge students' progress and adjust their teaching objectives in response to individual circumstances, potentially diminishing learning effectiveness [3]. A survey highlights that students' struggle to maintain focus and motivation during online learning is a significant reason why they may fall short of their expectations in micro-class learning [4]. Academic procrastination is prevalent among college students, with data analysis revealing conspicuous procrastination behavior in online learning [5]. It is crucial for learners to consciously cultivate self-control and reinforce self-awareness to successfully complete learning tasks [6]. Often, learners grapple with self-regulating their learning and may require both self-discipline and external support [7].

Interaction is among the most vital aspects for learners to acquire knowledge. In the context of learning, interaction encompasses activities involving engagement with teachers,

peers, learning materials, and the learning environment. The absence of face-to-face interaction in online courses presents a significant challenge, as students lack the opportunity for immediate feedback and direct communication with instructors when they encounter difficulties in their studies [8]. While online learning offers convenience, it cannot fully replace the value of in-person emotional communication, which, to some extent, impacts both student learning efficiency and teaching quality [9]. E-learning is associated with academic challenges, and the absence of face-to-face educator-student interaction is a pivotal factor, leading students to perceive traditional methods as superior [10],[11].

Learners express a preference for interactive microcontent, exercises, and immediate automated feedback as beneficial design elements for mobile micro-courses. Online learning often lacks collaborative opportunities, hindering students from developing teamwork and deep thinking skills [12]. Virtual classrooms may not engage students as effectively, and the dearth of traditional in-class social interactions results in limited real-time exchange of ideas among classmates [13]. Some students argue that distance education hampers group projects and collaborative work due to the lack of direct interaction [14].

Many micro-course resources remain unchanged for extended periods, making it challenging to tailor them to the ever-changing needs of students in today's fast-paced knowledge landscape. This static nature can lead to students losing interest in studying and even abandoning their pursuit of knowledge [15]. The inundation of negative content on the internet can also influence individuals' worldviews, life perspectives, and morals, leading to frustration and discontinuation of online learning [16]. Quality issues such as video playback problems, low resolution, and unclear content can quickly drain learners' motivation during the learning process, obstructing the widespread adoption of mobile learning and learner retention [17].

Key concerns for students revolve around the quality and quantity of mobile learning resources. The lack of specificity in course materials and issues with the quality of online learning resources significantly impact the retention rates of students in online courses. It is desirable for authoritative institutions to curate and provide high-quality resources that align with educational plans and learner needs, particularly in the context of extensive and complex teaching resources. Hence, this study investigates the impact of a novel segmentation process on enhancing the microlearning course structure within a blended learning environment. This approach integrates the Micro Learning technique with the course content of an IT module, specifically the "Information System" module, known for its practical components. The primary objectives of this research were as follows: (i) to evaluate the effectiveness of microlearning in IT education, (ii) to assess the impact of improved microlearning course structures on knowledge retention, and to determine the influence of segmentation strategies on knowledge retention.

The evolution of modern communication technology and the introduction of the "microlearning" theory have effectively bridged the gap between knowledge and learners, liberating them from the constraints of time and

space. It is expected that breaking instructional material into manageable, bite-sized portions will enhance motivation for learning and subsequently elevate learner engagement [19]. "Microlearning" represents an innovative approach to learning that aligns well with society's demands for lifelong education. It leverages communication technology to enable two-way communication, facilitating learning at any time and from any location. Typically, it employs succinct, brief content blocks to deliver and organize learning materials, emphasizing concise, loosely connected knowledge segments or modules within a limited timeframe. Microlearning primarily employs network communication technology to deliver content and facilitate interactions [20].

The distinctive features and significance of "microlearning" are readily apparent. It is typically centered around a single learning topic, featuring concise content chunks that usually span 5 to 15 minutes. This design caters to the human brain's attention span and includes time limits to prevent cognitive overload. The focus is on a clear and specific theme, which aids learners in clarifying their learning objectives and the problems they aim to solve. By concentrating on specific learning content, students can enhance their learning efficiency. Additionally, "microlearning" offers greater flexibility, as it eliminates traditional constraints related to location and time. Learners can engage in online learning activities according to their personal schedules, bolstering their enthusiasm and initiative for learning [21].

"Microlearning" can serve various purposes, such as being implemented as part of a flipped classroom approach, where learners complete microlearning activities before or after classroom teaching to reinforce knowledge concepts. Moreover, it serves as a versatile means of quickly delivering learning content and also supports social interaction, allowing learners to continuously enrich themselves and engage in self-development [22]. In the context of "microlearning," learners are not merely consumers of content; they can become content organizers. Learners with diverse interests and knowledge backgrounds can promote knowledge integration through interactive communication, forming a limitless repository of new knowledge [24].

Furthermore, "microlearning" can leverage online education tools like Moodle to facilitate teaching and communication interactions between educators and learners, even enabling educators to be present online. From a psychological perspective, "microlearning" can mitigate uncomfortable situations that some introverted or shy learners may encounter in traditional face-to-face classrooms, providing a more comfortable and accommodating learning environment [23].

Therefore, this research aims to investigate the impact of a proposed segmentation process implemented in content-based design on students' knowledge retention. The hypotheses are as follows:

H01: There is a statistically significant positive difference between the pre-test and post-test scores for students who engage in microlearning with the segmentation process.

H02: There is a statistically significant positive difference in knowledge retention gain scores between students who participate in microlearning with the segmentation process and those who do not.

3. METHODS

This study investigates the influence of a segmentation process on microlearning courses in a fully online learning environment. Specifically, it examines the effects of implementing the proposed segmentation process strategies in the course content of an IT subject, the Information System module, which focuses on practical activities. The research employs a pretest-posttest control group design to assess the effectiveness of these interventions. The pretest, measuring knowledge retention, was administered during the fourth week of the semester, while the posttest took place in the fourteenth week.

The research flow is illustrated in Figure 1. The study involved a total of 90 participants, divided into two groups. Group 'A' constituted the experimental group, which engaged with a module designed using the segmentation process, while Group 'B' served as the control group, participating in a microlearning module without integration of the segmentation process. The treatment condition was introduced to evaluate differences in outcomes. Both groups experienced the same peer learning environment, timeframe, and multimedia content but with varying organization of learning materials and knowledge structures.

To assess the effectiveness of the treatments, a pre- and post-test design was employed. One group, consisting of 45 students, underwent the treatment, while the other group, the control group of 45 students, did not receive any treatment but underwent the same testing procedures. Both groups were instructed to respond to a total of 3 short-answer questions and 2 essay questions on both the pre-test and post-test.

Quantitative research methodologies were employed to analyze the impact of the suggested segmentation procedure in microlearning on information retention and course completion throughout the semester-long lesson plan. The final exam scores were compared to the midterm exam grades to evaluate the results. The findings indicate that students demonstrate the capacity for self-directed learning within a self-regulated learning environment when utilizing the proposed segmentation approach.

The experimental groups underwent the pre-test for the two dependent variables, and the treatments were administered to these groups as they were relevant to the experimental process at that time. This design was chosen because it allows the researcher to assess participants' information retention and activity completion both before and after the experimental manipulations, facilitating a comparison of initial findings with the experiment's results. To account for potential absences and participant attrition during the experiment, 45 students were recruited for each treatment group, and the same student groups participated in both the pre-test and post-test assessments.

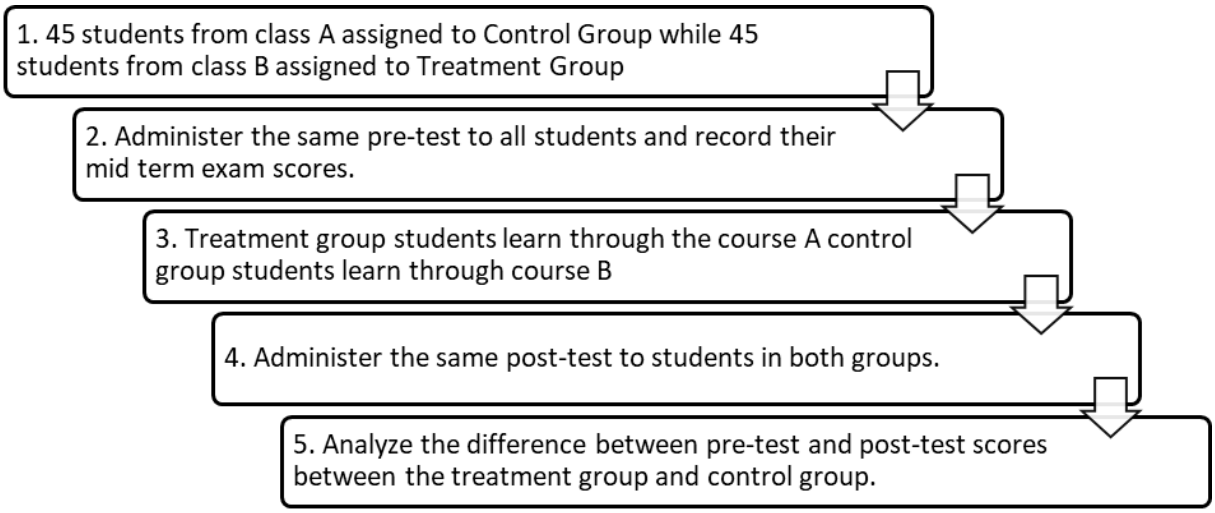

Figure 1. Research Flow

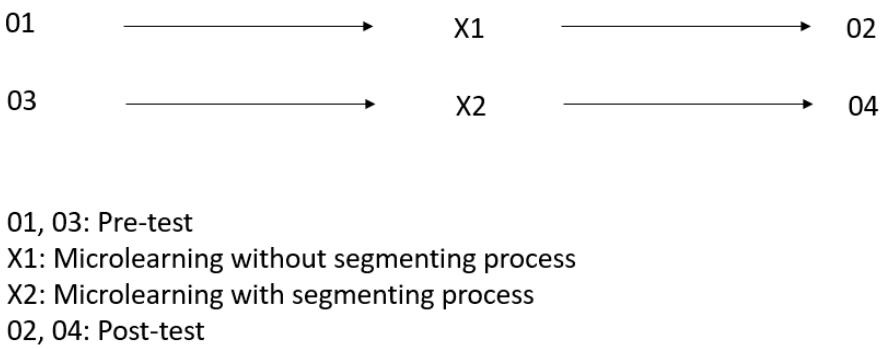

Figure 2. Design Notation

3. RESULTS AND DISCUSSION

H01: There is a positive significance difference between the pre-test and post-test scores for students who learn through microlearning with the segmenting process.

As indicated by the results shown in Table 1, the average score on the post-test was approximately 19 points higher compared to the pre-test. Furthermore, the Pearson correlation analysis between the pre-test and post-test scores (as detailed in Table 2) demonstrates a moderately positive correlation, with a coefficient value of $r = 0.259$ and a significance level of $p<0.05$. This suggests that, following their engagement with microlearning integrated with the segmentation process, the post-test results reveal an enhancement in students' knowledge retention compared to their pre-test performance.

Table 1. Paired samples statistics: Means

Paired samples statistics				
	Mean	N	Std. Deviation	Std. Error Mean
pre-test	61.22	90	15.214	1.604
post-test	79.89	90	13.474	1.420

Table 2. Paired samples statistics: Correlation

Correlations			
		pre-test	post-test
pre-test	Pearson Correlation	1	0.259*
	Sig. (2-tailed)		0.014
	N	90	90
post-test	Pearson Correlation	0.259*	1
	Sig. (2-tailed)	0.014	
	N	90	90

*Correlation is significant at the 0.05 level (2-tailed).

H02: There is a positive significance difference between gain scores of knowledge retention for students' that learning through microlearning with segmenting process & without segmenting process.

The critical value was more than 0.05, which is 0.357, as shown in the following table (Table 3). Therefore, the null hypothesis (H02) is rejected, which also suggests that there were no substantial differences between the pre-test and the post-test.

Table 3. ANOVA

ANOVA					
gain_scores					
	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	263.511	1	263.511	0.857	0.357
Within Groups	27048.489	88	307.369		
Total	27312.000	89			

4. CONCLUSION

The purpose of this study was to examine the relationship between 'before' (early in the To assess student knowledge retention in an Information System course at the end of the semester, we conducted examinations both before and after the course. The examination comprised three sections, including three short-answer questions and two essay questions. Notably, there was a significant difference in gain scores between the "before" and "post" tests, favoring those who underwent microlearning with the segmentation process over those who did not. Addressing real-world challenges related to Transaction Processing System components and networking and topology components necessitates a sound understanding of the relevant knowledge and procedures. The results of the examination revealed that students were able to address questions more directly and with greater clarity. It's worth noting that some students consistently made the same error on both the pre-test and post-test, possibly due to confusion between topology and its functions. The second component of the examination involved explaining a Decision Support Management System and its procedures. Students were required to provide a scenario and an example of a decision support system, as well as describe the role of such a system. The percentage of students providing correct responses remained consistently above 50%. Comparing pre-test and post-test results for each question showed an increase in the proportion of correct answers. This suggests that as the semester progressed, students developed a deeper conceptual understanding of the subject matter. This also indicates that students' ability to elucidate scenarios and processes related to information systems improved. Concerning essay questions about knowledge management systems, the results indicated varying significance for each question, with higher post-test percentages. This can be attributed to two primary factors. Firstly, the segmented nature of microlearning promotes sequential learning. The instructional materials were structured in a sequential manner, which aids students in identifying key terms and navigating the course content systematically. Secondly, the design of the learning resources distinguishes them from other online resources. For instance, lectures adhere to established pedagogical patterns, starting with introductions, followed by multiple subtopics, and concluding with summaries. Consequently, students grasp the relationships between information system components, such as servers, databases, and networking, in a visual manner, facilitating their comprehension of concepts and knowledge.

REFERENCES

[1] Pitman, T. & Broomhall, S. (2009). Australian universities, generic skills and lifelong learning, *International Journal of Lifelong Education*, 28:4, 439-458, DOI: 10.1080/02601370903031280.

[2] Aspin, D.N., & Chapman, J.D. (2000). Lifelong learning: concepts and conceptions. *International Journal of Lifelong Education*, 19:1, 2-19, DOI: 10.1080/026013700293421.

[3] De Gagne, J.C., Park, H.K., Hall, K., Woodward, A., Yamane, S., & Kim. S.S.

(2019). Microlearning in Health Professions Education: Scoping Review. *JMIR Med Educ*. 2019 Jul 23;5(2):e13997. doi: 10.2196/13997. PMID: 31339105; PMCID: PMC6683654.

[4] Zühal, S. & Hafize, K. (2021). The Effect of Mathematics Teaching Through Micro Learning in the ELearning Environment on Conceptual and Procedural Knowledge.

[5] Mary, D. & Joel, R. (2020). Microlearning: A New Learning Model. *Journal of Hospitality & Tourism Research*. 44. 109634802090157. 10.1177/1096348020901579.

[6] Cai, W., & Chen, Q. (2018). An Experimental Research of Augmented Reality Technology from the Perspective of Mobile Learning. 2018 *IEEE International Conference on Teaching, Assessment, and Learning for Engineering (TALE)*, 912–915.

[7] Renée, J. & van Leeuwen, Anouschka & Jeroen, J. & Liesbeth, K. (2020). A mixed method approach to studying self-regulated learning in MOOCs. *Frontline Learning Research*, *8,* 35-64. 10.14786/flr.v8i2.539.

[8] Jindong, W. & Yiqiang, C. & Shuji, H. & Feng, W. & Zhiqi, S. (2017). *Balanced Distribution Adaptation for Transfer Learning*. 1129-1134. 10.1109/ICDM.2017.150.

[9] Dingler, T., Weber, D., Pielot, M., Cooper, J., Chang, C.C., & Henze, N. (2017). Language learning on-the-go: opportune moments and design of mobile microlearning sessions. *Proceedings of the 19th International Conference on Human-Computer Interaction with Mobile Devices and Services*, 1–12.

[10] Ubaid, F., Saiful, A. & Lodya, S. (2021). *The Evaluation of E-learning Programs in Higher Education Using the CIPP Model* (The Empiric Study in Two Institutions of Sasmita Jaya Foundation). 10.4108/eai.17-7-2020.2303050.

[11] Azat, M., Norair, A., Aelita, S., Azat, G., Gulchachak, G., & Albina, G. (2021). Students' attitude to e-learning. *SHS Web of Conferences*. 97. 01042. 10.1051/shsconf/20219701042.

[12] Lee, Y.M. (2021) Mobile microlearning: a systematic literature review and its implications. *Interactive Learning Environments*, DOI: 10.1080/10494820.2021.1977964.

[13] Britt, R. (2006). Online education: a survey of faculty and students. *Radiologic technology*, *77*(3), 183-190.

[14] Kainat, A. & Muhammad, A. (2020). Online learning amid the COVID-19 pandemic: Students perspectives. *Journal of Pedagogical Research*. 1. 45-51. 10.33902/JPSP.2020261309.

[15] Erwen, Z., & Wenming, Z. (2017). Construction and Application of MOOC-based College English Micro Lesson System. *International Journal of Emerging Technologies in Learning (iJET)*, 12(02), 155–165.

[16] Emtinan, A. (2017). Microlearning: A Pedagogical Approach For Technology Integration.

[17] Tabares, M.S., Vallejo, P., Montoya, A. (2022). A feedback model applied in a ubiquitous microlearning environment using SECA rules. *J. Comput High Educ* (2022). https://doi.org/10.1007/s12528-021-09306-x.

[18] Tomas, J. & Radim, P. (2019). *Comparing the Effectiveness of Microlearning and eLearning Courses in the Education of Future Teachers*. 10.1109/ICETA48886.2019.9040034.

[19] Nikou, S. A., & Economides, A. A. (2018). Mobile-based assessment: A literature review of publications in major referred journals from 2009 to 2018. *Computers & Education, 125,* 101-119. https://doi.org/10.1016/j.compedu.2018.06.006

[20] Alqurashi, E. (2017). Microlearning: A pedagogical approach for technology integration. *The Turkish Online Journal of Educational Technology*, *16*, 942-947.

[21] Chong, S., Chua, F., & Lim, T.Y. (2021). Personalized Microlearning Resources Generation and Delivery: A Framework Design.

[22] Dixit, R., Yalagi, P.S., & Nirgude, M.A. (2021). Breaking the walls of classroom through Micro learning : Short burst of learning. *Journal of Physics: Conference Series, 1854*.

[23] Luminiţa, G. (2017). Microlearning an Evolving Elearning Trend. *Scientific Bulletin*. 22. 10.1515/bsaft-2017-0003.

[24] Max, B., Jan, R., Tobias, R., & Christoph, M. (2019). *From MOOCs to Micro Learning Activities*. 280-288. 10.1109/EDUCON.2019.8725043.