BILINGUAL PUBLISHING GROUP

Forum for Linguistic Studies

https://journals.bilpubgroup.com/index.php/fls

ARTICLE

On Determining Multiple Languages through Technological Examination for Conservation Management Using Machine Learning

Nguyen Minh Tuan [1] ***, Phan Thi Thanh Thuy*** [2*] ***, Ha Huy Nguyen Cuong*** [3] ***, Nguyen Trong Hien*** [4]

[1] *Faculty of Information Technology, Posts and Telecommunications Institute of Technology, 11 Nguyen Dinh Chieu, Dakao Ward, 1 District, Ho Chi Minh City 700000, Viet Nam*

[2] *University of Foreign Language Studies, The University of Danang (UFLS), Le Duan, Hai Chau, Da Nang 50000, Viet Nam*

[3] *Software Development Center, University of Danang, Le Duan, Hai Chau, Da Nang 50000, Viet Nam*

[4] *Faculty of Public Health, Pham Ngoc Thach University of Medicine, 2 Duong Quang Trung, 10 District, Ho Chi Minh City 700000, Viet Nam*

ABSTRACT

This study delves into the intricate linguistic history of Quang Nam province, a region rich in cultural and linguistic layers that existed long time ago. By analyzing place names and linguistic strata, the research uncovers traces of Mon Khmer and Austronesian languages, highlighting the region's deep historical connections and multilingualism within Cham Pa society. To further explore these linguistic complexities, we employed an interdisciplinary approach that integrates insights from comparative linguistics, archaeology, and cultural studies. A key contribution of this research is the application of advanced information technology, specifically data mining and artificial intelligence, to the preservation and analysis of this linguistic heritage. Using the YOLO-v8 deep learning model, we developed a system capable of accurately recognizing and classifying handwritten place names. The YOLO-v8 model, renowned for its powerful object detection capabilities, was instrumental in automating the analysis of large datasets, achieving high levels of accuracy in the recognition of diverse linguistic characters. This integration of AI not only enhances our ability to study historical language use but also provides an efficient and scalable solution for archiving and managing valuable cultural data. The results of this study contribute to both the

*CORRESPONDING AUTHOR:
Phan Thi Thanh Thuy, University of Foreign Language Studies, The University of Danang (UFLS), Le Duan, Hai Chau, Da Nang 50000, Viet Nam; Email: ptthuy84@ufl.udn.vn or Phanthithanhthuy1983@gmail.com

ARTICLE INFO
Received: 16 March 2025 | Revised: 18 April 2025 | Accepted: 23 April 2025 | Published Online: 8 May 2025
DOI: https://doi.org/10.30564/fls.v7i5.9110

CITATION
Tuan, N.M., Thuy, P.T.T., Cuong, H.H.N., et al., 2025. On Determining Multiple Languages through Technological Examination for Conservation Management Using Machine Learning. Forum for Linguistic Studies. 7(5): 643–654. DOI: https://doi.org/10.30564/fls.v7i5.9110

field of linguistic conservation and the development of AI-based tools for heritage preservation, ensuring the longevity and accessibility of the rich linguistic diversity of the region for future research.

Keywords: Place Name; Origin; Champa; Multilingual; Quang Nam Province; AI; Dataset

1. Introduction

Vietnam is a multilingual and multi-ethnic nation officially comprising 54 recognized ethnic groups, including the majority Kinh population and 53 ethnic minority communities. From a linguistic perspective, however, Tran Tri Doi identified 52 ethnic minority languages in Vietnam [1]. Among these, certain languages exhibit unique sociolinguistic characteristics. For instance, the O Du language spoken by the O Du people residing in western Nghe An no longer functions as a living language in everyday community life. Conversely, the Tay-Nung language is shared by two distinct ethnic groups: the Tay and the Nung. Ethnic minority populations in Vietnam often live both in dispersed and intermingled settlements, a demographic pattern that complicates the classification and enumeration of ethnic minority languages and also influences broader aspects of social development. This linguistic complexity is rooted, in part, in historical processes of migration and interethnic language contact.

In the contemporary context, Quang Nam province is home to a diverse range of ethnic groups, including the Co Tu, Ca Dong, Xe Dang, Gie-Trieng, and Hre, alongside the dominant Kinh (Viet) population. Historically, Quang Nam was also a key region of the ancient Champa civilization, and archaeological evidence continues to reveal significant cultural and historical vestiges left by the Cham people. To explore the extent of multilingualism in Champa society, we have selected a site in Quang Nam for interdisciplinary investigation. This research seeks to understand the historical and contemporary linguistic landscape of the region, shaped by centuries of sociocultural interaction. While toponymic studies have seen extensive development in Europe, such research remains relatively nascent in Vietnam, particularly studies that employ an interdisciplinary methodology with linguistics at the core. Numerous archaeological findings, including those from the Bau Du site in Quang Nam dating back approximately 10,000 years to the early Neolithic period, attest to the region's longstanding human habitation. To begin addressing the question of multilingualism in the Champa context, our initial efforts focus on analyzing place names of Champa and Mon–Khmer origin. Given that hydronyms (river names) and toponyms (village names) often retain ancient linguistic elements, we utilized comparative methods to determine their linguistic affiliations. Following the methodological framework proposed by Tran Tri Doi, we undertook a twofold process: comparing basic vocabulary as a necessary condition and identifying systematic phonetic correspondences as a sufficient condition. According to Doi, a language can only be definitively classified within a language family if both criteria are simultaneously satisfied [2].

To date, the historical study of the Vietnamese language has progressed significantly, with scholars relying primarily on the diachronic linguistic approach to uncover its origins and development. In our investigation of multilingualism within Champa society, we similarly adopt this diachronic perspective, which has proven to be essential in tracing linguistic transformations over time. While numerous researchers have contributed to the understanding of Vietnamese language history, we find the interpretations presented by Tran Tri Doi to be particularly compelling. From a diachronic standpoint, Tran Tri Doi argues that the Vietnamese language originated from the common Viet-Muong language family, specifically within the Mon-Khmer branch of the Austroasiatic family of Vietnamese. To such ends, "Vietnamese Language History: A Contribution to Understanding Vietnamese Culture," Doi emphasizes the importance of distinguishing between geographical Southeast Asia and historical (or cultural) Southeast Asia. He notes that although the Southeast Asian region is defined geographically, its cultures and heritage have widely evolved across multiple linguistic families. As we understand from his perspective, this complexity and interconnection provide a multi-layered view of the diverse and complex forces at play in language change in the region.

In line with Doi's perspective, we recognize Vietnamese are also referred to as part of the Viet–Muong group fundamentally Austroasiatic. For the purposes of this study, we deliberately set aside alternative hypotheses that attribute significant Chinese lexical derivations to Vietnamese or that propose non-Austroasiatic origins. These include Theurel's Austronesian hypothesis, as well as the positions of scholars such as Binh Nguyen Loc and Ho Le, who advocate for Vietnamese–Malay or Tai affiliations, drawing on earlier suggestions by H. Maspero. Through critical engagement with these diverse viewpoints spanning the work of scholars like M.P., A. Hauricourt, and M. Avest we gain a broader understanding of the complex linguistic landscape of Southeast Asia. These varying arguments illuminate the contested nature of language classification in the region and underscore the importance of a multi-perspective, historically grounded approach in examining the evolution of the Vietnamese language.

2. Related Work

The Results of Archaeological and Historical Research Initially Identified Language Classes

Archaeological, historical, and cultural research—carried out by notable scholars such as Ho Xuan Tinh, Vu Cong Quy, Huynh Cong Ba, and Tran Quoc Vuong—provides compelling evidence that the region of Quang Nam has been continuously inhabited since ancient times. A significant discovery in 1981 at the Bau Du archaeological site unearthed five burials containing human remains, offering clear proof of early human settlement in the area [1,3]. Based on these findings, scholars have posited that prehistoric human activity in Quang Nam dates back six to seven thousand years, with evidence of subsistence practices such as hunting, fishing, shellfish gathering, and notably, early agricultural activity, including the cultivation of crops like water potatoes and yams [2,4]. Further archaeological excavations in locations such as Nui Thanh and Dai Loc have revealed numerous jar burials, many of which contained artifacts made from bronze and iron. These findings are attributed to the Sa Huynh culture, a significant prehistoric culture along Vietnam's Central Coast. Analysis of material remains has led researchers to suggest that the Sa Huynh culture may have served as a direct cultural predecessor to the ancient Champa civilization. This hypothesis is supported by archaeological evidence indicating a cultural continuity from the Sa Huynh period to the emergence of the Champa state, suggesting that ancient Champa inhabitants were likely descendants of late Sa Huynh communities.

The diversity in the materials used for jar burials, including agate, iron, and ceramics, is interpreted by Huynh Cong Ba as indicative of early forms of social stratification. This differentiation is seen as a precursor to the emergence of a state-level society. Furthermore, specific burial artifacts such as a Dai Lanh jar dating to a period close to the establishment of the Northern Cham State by Khu Lien in the 2nd century CE, and a Cam Ha jar dated to approximately the 1st century BCE to 1st century CE lend additional support to the argument that sociopolitical organization was already taking shape prior to the official founding of the Champa state. Importantly, this body of evidence suggests that prior to the rise of Champa, Quang Nam was already home to a diverse array of ethnic communities, evolving along different cultural trajectories. Linguistic studies further reinforce this claim. According to Ha Van Tan, and based on calculations by historical linguists G. Diffloth and I. Peiros, several Mon-Khmer languages the subgroup to which Vietnamese belongs began to diverge around 4000 BCE and 4200 BCE, respectively. These findings indicate that the Austroasiatic language family from which many of the region's languages are derived has deep indigenous roots in the area. Thus, both archaeological and linguistic evidence point to the long standing presence of Austroasiatic speaking communities in Quang Nam, long before the formation of the Champa polity [5,6].

In a corresponding historical timeframe, Professor Tran Tri Doi offers an insightful linguistic interpretation concerning the Lac Viet ethnonym and the language of the Dong Son culture. In his analysis, he contends that from the perspective of Vietnamese historical linguistics, there is little conclusive evidence to disprove the possibility that the inhabitants of Dong Son culture considered a foundational civilization in ancient Vietnam were indeed indigenous peoples who spoke a pre-Vietnamese language. Doi refers to this speech community as part of the core geographical region, or the homeland, of Austroasiatic languages, thereby reinforcing the notion of long-term linguistic continuity in the region [7]. Drawing from archaeological, historical, and cultural research, it becomes evident that Quang Nam possesses a rich stratification of linguistic layers. Initially shaped by indigenous Austroasiatic languages, this region later witnessed significant linguistic and cultural transformation following the emergence of the Champa state. As part of its sociopolitical development, the Champa civilization established a formal Cham script system to encode cultural, religious, and administrative information.

According to Po Dharma, this classical Cham script primarily used in inscriptions was developed from Indian Sanskrit characters, yet it underwent significant modifications to align with the cultural and aesthetic preferences of the Champa kingdom. This script served as a prominent medium of expression from the 4th to the 15th centuries, marking a flourishing period of written cultural production in Champa. However, Po Dharma also notes that classical Cham has since fallen into disuse, and at this point, no living Cham individuals possess full understanding of this classical form [3]. Thus, the linguistic history of Quang Nam reflects a complex evolution beginning with native Austroasiatic roots, followed by the development of a distinctive Cham literary tradition, and marked by successive cultural overlays. This layered history highlights the region as a significant locus of linguistic and cultural convergence in early Southeast Asian civilization [3]. The image of the Yen Chau bronze stele is shown in **Figure 1**.

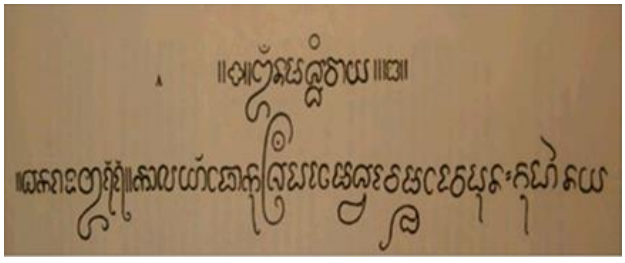

Figure 1. The still-hot image of the cham people's identification word when we went to field as field farmers in Quang Nam Province.

These findings also support the inference that during this historical period, residents of the region were likely bilingual, utilizing both the Cham language and languages from the Mon-Khmer linguistic bloc, as assessed by Tran Tri Doi [1]. Doi asserts that, prior to the territorial expansion of Dai Viet, this land was under the administration of the Champa kingdom. Therefore, it is reasonable to conclude that the communities inhabiting the area spoke Cham or possibly Mon-Khmer languages as their native tongues. With the annexation of this region into the Dai Viet state, where Vietnamese became the national language, it was inevitable that residents had to adopt Vietnamese as the dominant medium of communication in public and social life [8]. Despite centuries of transformation, we maintain that Quang Nam today still preserves numerous place names of Cham origin. However, several scholars argue that many of these names have become obscured over time due to linguistic overlay, particularly by Sino-Vietnamese influences, making them more challenging to decipher and analyze. Place names such as Cầu Nhi, Trà Kiệu, and Trà Nhiêu are cited as examples of such transformations. Researchers emphasize the importance of methodological frameworks such as the identification of basic lexical classes and the reconstruction of phonological evolution through linguistic change rules as effective strategies for classifying Vietnamese within the Vietic group of the Mon-Khmer branch of the Austroasiatic language family. While Ha Van Tan posits that Vietnamese originates from the common Viet-Muong language group, Tran Tri Doi refines this by identifying it as Proto-Viet–Muong (or Proto-Vietic) [9].

In a parallel thread of linguistic and technological exploration, there has been increasing attention toward the application of artificial intelligence (AI) in handwriting recognition, particularly for Han and Chinese characters. This field has seen rapid advancements in recent years, largely due to the proliferation of deep learning methods, especially convolutional neural networks (CNNs), which have significantly enhanced recognition accuracy for complex character systems. This trajectory reflects a broader shift in the field of optical character recognition (OCR) from traditional machine learning approaches to more sophisticated deep learning architectures [10,11]. A comprehensive benchmark study conducted offered a foundational comparison of handwritten Chinese character recognition online and offline, while subsequent work introduced a neural network-based recognition system capable of adapting to varied styles and complexities of handwriting [11,12]. This was further advanced by Wu and Zhong, who demonstrated the superiority of CNNs in handling the intricacies of handwritten Chinese scripts [12–14]. The authors also proposed a refined CNN architecture for improved generalization across diverse handwriting styles, while optimizing CNN models for embedded systems, addressing constraints in processing power and memory [15]. In particular, DropDist, a CNN-based model tailored for large-scale handwritten character recognition using smaller datasets, was introduced as a key innovation for applications with limited training data [16].

Further breakthroughs include Zhang, who integrated CNNs with Connectionist Temporal Classification (CTC) to jointly manage segmentation and recognition, and Zhao, who employed a CNN-LSTM framework to harness bidirectional stroke sequence learning, thus improving recognition accuracy by incorporating temporal dependencies in character writing [17,18]. Recent contributions from transfer learning explored the use of pre-trained models, effectively improving performance while reducing training overhead [19]. Building on this, Chen and Zhang proposed a highly efficient CNN-based architecture that optimized computational performance without sacrificing recognition accuracy, marking a new milestone in offline handwriting recognition for Chinese characters [20,21]. Together, these advancements underscore the evolution of handwriting recognition technologies from early algorithmic approaches to state-of-the-art deep learning models, setting the stage for further innovations in multilingual character recognition and digital preservation of linguistic heritage.

3. Dataset

This study introduces a dataset compiled during extensive field surveys conducted in Quang Nam province, which holds considerable scientific value. As a repository of cultural and linguistic heritage, this dataset is not only a rich source of information but also represents a critical resource that warrants digitization and long-term preservation for future generations. The dataset is particularly comprehensive, encompassing four columns and approximately 50,000 entries, as illustrated in **Tables 1** and **2**. It categorizes place names into three main groups: “General” names, “Proper Names,” and “Han-Nom characters.” This structured database provides a robust foundation for further linguistic, historical, and cultural analysis, especially within the interdisciplinary context of toponymy and language contact studies.

Table 1. Place name of an administrative unit in Dien ban district, Quang Nam Province.

Địa danh Chăm Pa cổ	Ancient Champa Landmark	Ghi chú (Notes)
Simhapura (Sư tử thành)	Simhapura (Lion City)	First capital of Champa, now in Tra Kieu (Duy Xuyen District)
Amaravati	Amaravati	A major Champa state, now in Quang Nam region
Mỹ Sơn (Mỹ Sơn Thánh địa)	My Son (My Son Sanctuary)	The most significant complex of ancient Champa temples and towers, now in Duy Xuyen
Trà Kiệu	Tra Kieu	Ancient citadel and major religious center of Champa
Đồng Dương	Dong Duong	Champa Buddhist center in the 9th century, now in Thang Binh
Khu vực Chiêm Sơn	Chiem Son Area	Site with Champa temple and tower ruins in Duy Xuyen
An Mỹ	An My	Area with Champa tower ruins (near Tam Ky)
Hòn Tàu, Hòn Ông	Hon Tau, Hon Ong	Sacred mountain peaks with traces of Champa presence

Table 2. Place name of an administrative unit in Dien ban district, Quang Nam Province.

Place Name	Location	Chăm Pa Origin	Imagery Origin (Related to Chăm Script)
Bồng Miêu	Phú Ninh District	"Bôn" (hill) + "Mưh" (gold): "Gold hill".	Image of an ancient gold mining hill. No recorded Chăm script inscriptions, but Chăm artifacts may be found nearby (Quảng Nam Museum).
Trà Kiệu	Duy Sơn Commune, Duy Xuyên District	"Trà" from "Ia/Ya" (water, river), Indrapura capital.	Ruins of Simhapura capital with ancient Chăm script inscriptions (4th –5th century), such as Bhadravarman stele. Apsara reliefs and Sanskrit on stone at Đà Nẵng Museum.
Trà Na	Bắc Trà My District	"Trà" from "Ia/Ya" (water, river).	Mountainous rivers and streams. No specific Chăm script inscriptions recorded, but possibly linked to Sa Huỳnh culture (pre-Chăm Pa).
Trà Ngâm	Nam Trà My District	"Trà" from "Ia/Ya" (water, river).	Mountainous forest landscape. No recorded Chăm script inscriptions, but Chăm artifacts may be found at local museums.
Bà Dàng	Various areas in Quảng Nam	"Ian/Yan" (deity), related to worship.	Shrines or spiritual spaces. No recorded Chăm script inscriptions, but possibly linked to Chăm Pa worship sites.
Bồ Bồ	Bồ Bồ Mountain, Đại Lộc District	"Pô" (deity), related to beliefs.	Sacred mountain. No recorded Chăm script inscriptions, but Chăm artifacts may be found nearby (Quảng Nam Museum).
Bồ Mưng	Quế Phú Commune, Quế Sơn District	"Pô" (deity).	Rural village with spiritual traces. No recorded Chăm script inscriptions, but related artifacts may be at museums.
Thạch Bồ	Tam Xuân Commune, Núi Thành District	"Pô" (deity) + "Thạch" (stone).	Sacred stone mound. No recorded Chăm script inscriptions, but possibly linked to nearby Chăm sites.
Mỹ Sơn	Duy Phú Commune, Duy Xuyên District	"Mỹ" (beautiful, sacred) + "Sơn" (mountain), Chăm Pa holy site.	Chăm tower complex with ancient Chăm and Sanskrit inscriptions (7th –13th century), like Mỹ Sơn A1 stele, stored at Đà Nẵng Chăm Museum.
Đồng Dương	Bình Định Bắc Commune, Thăng Bình District	Name of Buddhist monastery, linked to King Indravarman II.	Buddhist monastery ruins with Chăm and Sanskrit inscriptions (9th century), Bodhisattva statues, and reliefs, stored at Đà Nẵng Chăm Museum.

4. Research Place Names in Quang Nam to Identify the Multilingual Status

Historically, the territory of present-day Quang Nam was once under the governance of the Champa kingdom, which established two major capitals in the region: Simhapura and Indrapura. The recorded history of Champa begins in the 2nd century CE, and the earliest inhabitants of this region are believed to have been Mon-Khmer ethnic groups. From a historical and linguistic perspective, the formation of the Champa kingdom likely resulted from the consolidation of several smaller polities, commonly referred to in Southeast Asian political theory as "Mandala" systems. As the kingdom expanded and solidified its power, it exerted control over various subordinate groups, with the Cham language serving as the administrative and cultural

lingua franca. However, it is reasonable to infer that the Mon-Khmer language, known locally as Pylanh, continued to be spoken among lower social strata, suggesting that bilingualism or even multilingualism was a defining feature of Champa society. This specific type of linguistic diversity is reflected in the toponymic landscape of Quang Nam, where multiple linguistic strata overlap, revealing traces of different cultural and ethnic influences over time. Therefore, any attempt to identify and classify the multilingual characteristics of Champa society must be approached with caution, as the linguistic layers are deeply intertwined and shaped by centuries of contact, migration, and cultural integration.

4.1. Explanations About Place Names of Cham Origin

Research by Bui Trong Ngoan has revealed intriguing insights into the origins of major river names in Quang Nam, such as Thu Bồn, Ba Rén, and Vụ Gia, suggesting that these names retain a high degree of linguistic accuracy [9]. However, in the specific case of the Thu Bồn River, while the author explores multiple etymological possibilities ranging from "Bô Bô," "Pô Inu Nagar," to "Simhapura", he ultimately concludes that these names lack clear phonetic connections. Instead, he observes that the ritual traditions of the Thu Bồn village, notably the sacrifice to Lady Bô Bô (Po Inu Nagar), reflect cultural continuity from the ancient Champa civilization. The diversity in the names used in these practices—both divinity types (e.g., "god," "king," or "sin"), yet with phrases that the term "Thu Bồn" does not have obvious origins from either the Cham expressions. Rather, he suggests it might originate from the Cham word "tabbok", meaning "mound", likely referencing the local geography. From our standpoint, it is plausible that the Thu Bồn River name emerged during the Champa period, shaped by cultural practices, such as the matrilineal traditions of the Cham and their close relationship with riverine environments.

A similar multilingual heritage can be observed in the naming of the Cầu Nhi River, which may also reflect Cham linguistic influences. In the Cham language, the term "Kunýk" includes the pre-syllable "ku" (interpreted as "sentence") and the root "nýk" (meaning "gold," "yellow," or "turmeric"). The semantic field around "gold" evokes connotations of fertility and richness, leading to the interpretation of Cầu Nhi as a fertile, alluvial river. Another place name, Cầu Lau, is believed to stem from a Cham origin as well, functioning as a transliteration of "pulau" (island), adapted to Quang Nam phonology as "pulao." This suggests that phonetic evolution and local variation in linguistic borrowing have significantly influenced place name formations in the region.

Bui Trong Ngoan also notes that toponyms such as Trà Kiệu, Trà Nhiêu, and Cầu Nhi are often overlaid with Sino-Vietnamese phonological structures, making it difficult to clearly attribute their origins to either Cham or Mon-Khmer languages. For instance, Trà Kiệu village, established in the 17^{th} century, sits atop the ruins of Simhapura, the ancient capital of Champa (4^{th} – 8^{th} century), located in modern-day Duy Xuyên. This area likely retains deep linguistic and cultural residues from the Champa civilization.

Linguistic elements such as "Trà" are prevalent in the central region, e.g., Trà Khúc, Trà Bồng, Trà Câu, Trà Khe suggesting a systematic naming pattern. Some scholars link "Trà" to "Cham Cha," believed to be an ancestral figure of the Cham people. From a broader linguistic perspective, "Trà" might be a derivation of the Austroasiatic ya, which could mean "river", or alternatively, it could signify "village" or "region." This dual interpretation leads to the hypothesis that Trà Kiệu may denote a large settlement, while Trà Khúc or Trà Bồng may reference river systems named after physical features.

Drawing on the extensive research of Tran Tri Doi, we support the theory that the Vietnamese mainland was initially inhabited by indigenous Austroasiatic peoples, followed by early interactions with Austronesian groups. These groups, particularly under Indian cultural influence, produced early texts and inscriptions, including Vietnamese records dating back to the 2nd century CE in the Lâm Ấp polity. Prior to the expansion of the Đại Việt state, the southern regions of Vietnam were predominantly populated by speakers of Mon-Khmer and Austronesian languages, reflecting a multilayered cultural and linguistic substratum.

Consequently, the concept of "Mandala", commonly used in Southeast Asian studies to describe political configurations of loosely associated but culturally interconnected groups, is especially relevant to Champa society. In such a system, the dominant Champa state occupied the central role, while diverse ethnic communities, often retaining their Mon-Khmer traditions, existed within its periphery. This complex political and linguistic configuration provides vital context for understanding the stratified and hybrid toponymy observed throughout the Quang Nam region today.

4.2. Explanations About Place Names of Vietnamese Origin

Several river names in Quang Nam Province bear clear Vietnamese linguistic origins, such as Chợ Cui River, Bến Gia, Bến Giang, Cái, and Con. For instance, Chợ Cui River is known in Nom script and was formerly referred to as Sài Thị River, where "Sài" denotes firewood and "Thị" denotes market in classical Han characters. This river, also

referred to as Sài Thị Giang ("Giang" meaning "river"), is a downstream branch of the Thu Bồn River, adjacent to a market known as Cui Market. This toponym reflects a metonymic naming strategy, wherein the name of a prominent local feature, such as a market, is extended to label the river. From our perspective, the term "Sài Thị Giang" may have served as a Sino-Vietnamese formal designation, while the more vernacular and widely used Chợ Cui has persisted in common usage today. This evolution exemplifies the dual-layered naming practices formal versus colloquial that often characterize Vietnamese place names.

Another example is the Con River (also known as the Golden River), which originates from the former Hiên District and serves as a major tributary of the Vũ Gia River within the Thu Bồn river system, flowing through Đông Giang and Đại Lộc districts. Across various official and historical documents including the List of Intra-Provincial River Basins the name is inconsistently recorded as either Sông Côn or Sông Con. This discrepancy likely reflects phonological variations in the Quang Nam dialect, where the nasal ending "-on" is sometimes articulated as "-on" or "-ôn" due to local pronunciation features. Moreover, the appearance of "Kon" may derive from phonetic transcription practices, especially since Vietnamese Latin characters traditionally avoid the use of "k" in favor of "c" or "qu." Nonetheless, some transcriptions render the pronunciation as /kon1/, suggesting possible Austroasiatic or indigenous linguistic influences. This point remains a subject of ongoing investigation.

In the western mountainous regions bordering Laos, the presence of multiple ethnic groups, such as the Cơ Tu, Xơ Đăng, Cor, and Gié-Trieng, alongside ethnic Vietnamese populations, further reflects the area's rich cultural and linguistic diversity. Place names in these regions, such as Đắk Di, Đắk Mi, Đắk Sê, Đắk Pring Ta Vi, incorporate the element "Đắk," a Mon-Khmer term particularly from the northern Bahnaric subgroup which translates to "water" or "river." These names illustrate the deep-rooted influence of indigenous languages in the geographical nomenclature of the region.

Taken together, this evidence supports the view that Quang Nam represents a historically multilayered and multilingual cultural landscape. The accretion of Austroasiatic, Austronesian, Sino-Vietnamese, and indigenous linguistic features in the place names of rivers and settlements reflects a long process of cultural contact, adaptation, and integration, reinforcing the importance of toponymy in tracing historical and ethnolinguistic interactions in Central Vietnam.

4.3. YOLOv8 Algorithm for Chinese Character Recognition

Input:

+ Image: Input image containing Chinese characters (e.g., 640x640 pixels).
+ pre_trained_model: YOLOv8 model fine-tuned for Chinese character detection and recognition.
+ confidence_threshold: Minimum confidence score to keep a detection.
+ iou_threshold: IoU threshold to apply Non-Maximum Suppression (NMS).

Output:

+ A list of detected Chinese characters with:
· Bounding box coordinates (x, y, width, height).
· Recognized character label (class).
· Confidence score for each detection.

The construction is conducted as follows:

1) Step 1: Preprocess the input image
2) Step 2: Feature extraction
3) Step 3: Feature aggregation (Neck)
4) Step 4: Predictions (Head)

- "Pass feature" maps to the pre_trained_model.head to make predictions:
- "Bounding boxes" for each character: (x, y, width, height).
- "Class label" for each character (from a predefined set of Chinese characters).
- "Confidence score" for each detection.

5) Step 5: Filter predictions

- Discard bounding boxes with confidence score < confidence_threshold.
- For each remaining bounding box, assign the class label with the highest probability.

6) Step 6: Apply Non-Maximum Suppression (NMS)

- For overlapping bounding boxes:
- Compute IoU between boxes.
- Keep only the box with the highest confidence score if IoU > iou_threshold.
- Remove redundant boxes to avoid multiple detections of the same character.

7) Step 7: Output the results

- Return the list of detected Chinese characters, including:
- Bounding box coordinates.
- Recognized character class label.
- Confidence score of each character recognition.

The results of prediction are shown in **Figure 2**.

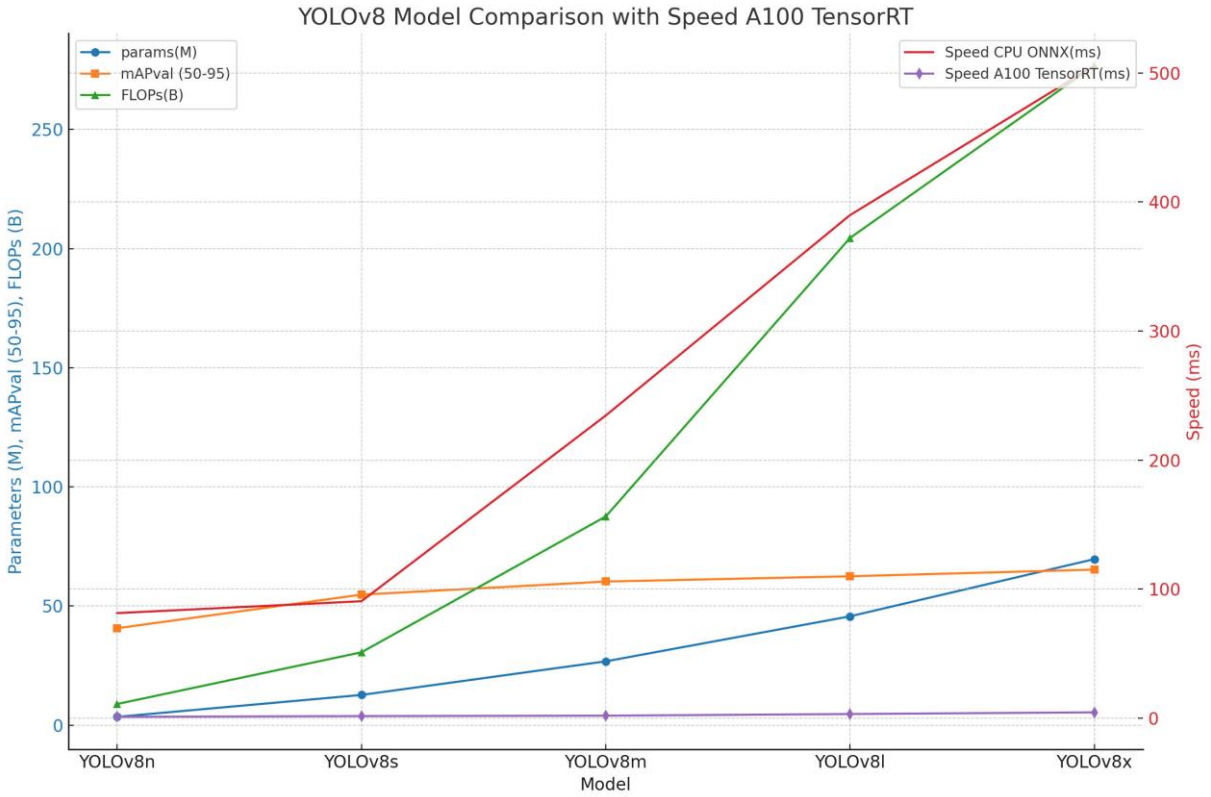

Figure 2. Performance of prediction.

5. Discussion

In this study, we examine a digitized dataset of place names collected from various localities within Duy Xuyên and Điện Bàn districts in Quảng Nam Province. The dataset includes entries gathered through field surveys and historical records, with each location represented by three key attributes: (1) a general classification under the attribute "Chung," referring to the type of administrative unit such as commune or village; (2) the proper name of the commune or village; and (3) the corresponding name in Hán-Nôm script, when available. These attributes provide a comprehensive linguistic and cultural representation of the local geography.

To optimize the analytical process, we apply correlation-based feature selection to reduce dimensionality, narrowing the dataset from an initial 10 attributes to 4 core features that carry the most predictive significance. This feature selection step not only simplifies the dataset but also enhances the efficiency and accuracy of subsequent computational models.

Our approach integrates techniques from natural language processing (NLP) and deep learning to address the challenges inherent in analyzing toponymic data. Specifically, we employ deep learning algorithms to predict missing place names, leveraging learned patterns in phonetic, semantic, and historical attributes. Additionally, unsupervised clustering techniques are utilized to group similar place names, enabling the identification of linguistic trends and geographic naming conventions over time.

Further, we explore AI-driven models to suggest potential place names grounded in historical context and linguistic patterns, contributing to the broader effort of cultural preservation and historical reconstruction. Overall, this work not only digitizes but also semantically enriches over 500 commune-level place names across the two districts, forming a valuable resource for linguists, historians, and computational researchers interested in Vietnam's toponymic landscape shown as "公社." The commune word images were captured from the village gate, which was experimentally manipulated using a fake "F" to simulate the environment in which the investigation took place.

This process of photographing and digitizing the word images posed significant challenges due to their inherent characteristics: the images were often blurred and difficult to distinguish, and the characters displayed ancient origins that made identification particularly complex. After organizing the training data for the deep learning model, we proceeded with three testing scenarios. In the first scenario, we utilized 88% of the dataset for training, 8% for validation, and 4% for testing, as illustrated in **Figure 3**. The results of this experimental setup yielded an impressive

recognition accuracy of 97.8% for the commune letters, as demonstrated in **Figures 4** and **5**.

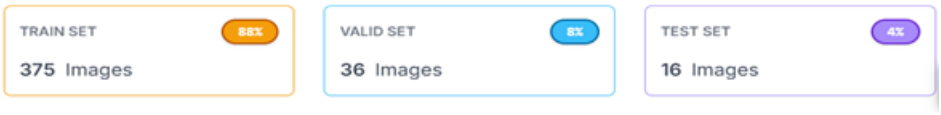

Figure 3. Dataset division into subsets.

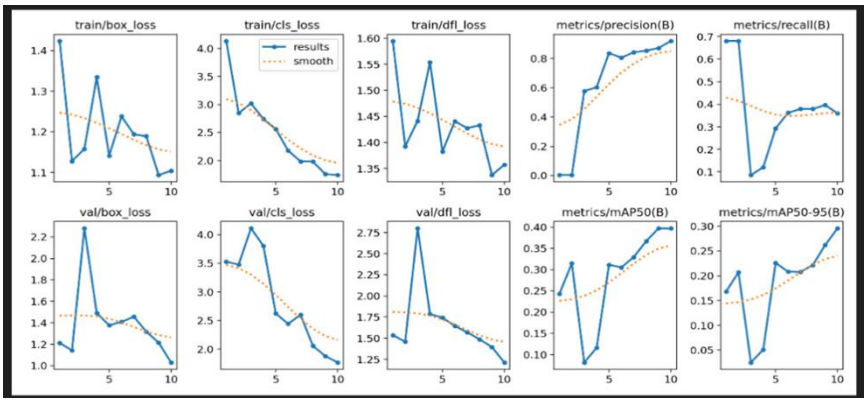

Figure 4. Results after training the YOLOv8 model.

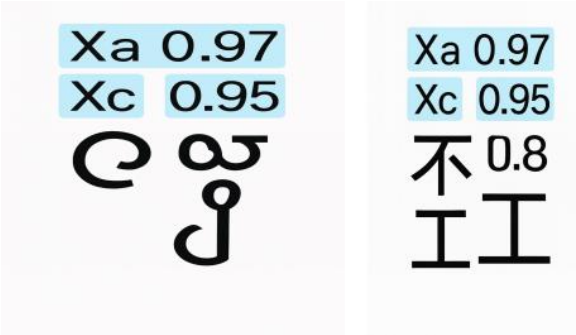

Figure 5. Image recognition results using YOLOv8.

The result is a robust dataset comprising over 50,000 entries, encompassing a broad spectrum of place names with a particular emphasis on common administrative designations such as “Commune” (Xã), “Village” (Thôn), and other related terms. The sheer volume of these 50,000 place words constitutes a comprehensive and significant collection. This large dataset not only captures the richness of local toponymy but also provides a detailed view of the linguistic diversity within Quang Nam Province. It thoroughly explores the region’s administrative structure and its influence on language use.

The dataset serves as a valuable resource for in-depth analysis of linguistic diversity within Champa society and other local communities in the province. By examining the frequency and distribution of place names, we can identify patterns in language use and pinpoint the most commonly spoken languages across different administrative areas. This highlights prevailing linguistic trends and offers insights into the vitality of languages in the region. Through geographical analysis, we gain a better understanding of the spatial distribution of languages and dialects, which is crucial for mapping language variations across Quang Nam Province.

Furthermore, the correlation between place names and linguistic or ethnic groups can shed light on the variations in language across different regions. This knowledge is especially valuable for identifying areas where certain languages are either dominant or endangered. The dataset thus plays a critical role in shaping language conservation strategies. Areas with high concentrations of endangered or less commonly spoken languages can be targeted for preservation efforts.

In addition to its conservation potential, the dataset can contribute to the development of educational materials and efforts to raise awareness about the region’s linguistic heritage. This helps preserve these languages for future generations. Place names, often rich with cultural and historical significance, offer a unique lens through which we can explore the relationship between language, culture, and history. Analyzing these names provides an opportunity to preserve the intangible cultural heritage embedded within them, strengthening the connection between language and its cultural roots.

In our study on character recognition for handwritten local language identifiers in Quang Nam Province, we adopted and improved several techniques previously explored by other research groups [11,12,16,17]. Our primary objective was to enhance model performance while addressing the unique challenges posed by the linguistic and ethnic differences across language groups in the region.

Model Modifications and Data Preprocessing:

(1) Image Enhancement: The ImageDataGenerator was optimized with parameters such as shear_range, zoom_range, and rotation_range to increase data diversity. Padding images to a fixed size helped

standardize input dimensions, leading to more consistent training. These modifications resulted in improved performance, with the model achieving a validation accuracy of 95% and a test accuracy of 84.2%. However, occasional mispredictions suggested the need for additional fine-tuning.

(2) Color Transformation and Background Noise: The conversion from RGB to grayscale simplified the input data and altered the model's behavior slightly. After optimizing the model with new parameters, a yellow background with added noise was introduced to simulate real-world conditions. These changes led to improved model performance, reaching a validation accuracy of 92% and a test accuracy of 84%.

Addressing Imbalanced Data:

Data imbalance emerged as a significant challenge, with the training dataset containing skewed character frequencies. Two techniques were tested to resolve this:

(1) Undersampling: This method involved removing images from overrepresented classes, but it risked losing important data variations and reducing dataset size.

(2) Oversampling: Replicating images from underrepresented classes proved to be more effective, as it maintained the diversity of the dataset and better addressed the imbalance. Ultimately, oversampling was chosen for the final Model.

Automated Model Selection (TPOT): To further optimize the model, we applied TPOT, an automated machine learning tool that fine-tunes model architecture and hyperparameters. This resulted in a significant improvement, with validation accuracy increasing to 97.6% (from 50,000 samples) and test accuracy reaching 89.5% (from 4,000 samples). This demonstrated a substantial boost in generalization performance.

We also explored the method of Chinese handwriting recognition using convolutional neural networks (CNN) and median filtering [21]. Comparing the results, we observed several differences in performance and efficiency:

(1) Median Filtering: Median filtering was applied to input images to reduce noise, and the images were resized to 28x28 pixels to meet the CNN's input requirements.

(2) Data Reading: All images were converted into matrices and split into training and testing sets.

(3) Convolutional Neural Network: The model consisted of two convolutional layers (with 6 and 12 filters), which produced 12 feature maps, each measuring 4x4.

(4) Fully Connected Layer: This layer made the final classification and determined the correct character recognition output.

(5) Training Process: The model was trained over several epochs, adjusting weights as needed. After 5000 epochs, the model achieved an accuracy of 97.28%.

(6) Challenges: Despite the high accuracy, the training time was considerable, with 5000 epochs taking 12,348 seconds. Additionally, the dataset size was relatively small, which could limit the generalization capacity of the model.

This paper presents a solution for mining data warehouses and advancing research in the application of artificial intelligence, machine learning, and deep learning to support handwritten place name character recognition using a convolutional neural network (CNN) model that has been deployed in many fields of applied science [22–25]. The proposed model, based on the YOLO-v8 architecture, demonstrates that after approximately 5,000 training epochs, it can achieve an accuracy of up to 98%. The enhanced YOLO-v8 model was also compared with a model utilizing random hyperparameters under the same conditions, including the same dataset and network architecture. The results indicate that the YOLO-v8 model outperforms the random hyperparameter model. However, to further improve classification accuracy, hyperparameter optimization remains a promising area of exploration for deep learning models. This method seeks to identify the optimal set of hyperparameters to enhance model performance, including improving classification accuracy and computational efficiency. Big data and smart data in open education and research, machine learning for open education and research, and personalized learning and research environment design are some of the subjects covered by the new technology, which focuses on cloud-based learning resources, platforms, and infrastructures [26,27]. This machine tools also provide new technology for establishing and analysis new data contributing in conserving the historical culture [28,29].

6. Conclusions

In conclusion, our research on the linguistic diversity of Quang Nam, combined with data mining and artificial intelligence techniques, has enabled us to discover significant information on the complex linguistic heritage of the region. By applying deep learning models, particularly the YOLO-v8 architecture, we have made substantial advancements in the automated recognition of handwritten place names. The use of the YOLO-v8 model has proven highly effective, demonstrating impressive performance in identifying and classifying characters with high accuracy due to its robust object detection capabilities.

The incorporation of data mining and artificial intelligence not only facilitates the automatic recognition of place names but also offers a scalable solution for processing large volumes of linguistic data. The YOLO-v8 deep learning model, with its enhanced accuracy and efficiency, has the potential to handle the complexities and variations in handwritten characters, especially those from

diverse language families like Cham and Mon-Khmer, as well as Vietnamese.

This approach opens new avenues for further research in both linguistic conservation and AI applications. By leveraging machine learning techniques and sophisticated models like YOLO-v8, future studies can refine these methodologies to achieve even greater accuracy and efficiency, thus enriching our understanding of the region's linguistic history. Moreover, this integration of technology in linguistic research provides a robust framework for the preservation and digital archiving of linguistic data, ensuring that this rich cultural heritage can be systematically conserved for future generations.

Author Contributions

Conceptualization, N.M.T., P.T.T.T., H.H.C.N., and N.T.H.; methodology, N.M.T., P.T.T.T., H.H.C.N., and N.T.H.; software, N.M.T. and P.T.T.T.; validation, P.T.T.T., H.H.C.N., and N.T.H.; formal analysis, H.H.C.N. and N.T.H.; investigation, N.M.T. and P.T.T.T.; resources, H.H.C.N. and N.T.H.; data curation, N.M.T. and P.T.T.T.; writing—original draft preparation, P.T.T.T.; writing—review and editing, P.T.T.T., H.H.C.N., and N.T.H.; visualization, N.M.T., P.T.T.T., H.H.C.N., and N.T.H.; supervision, H.H.C.N. and N.T.H. All authors have read and agreed to the published version of the manuscript.

Funding

This work was supported by Ministry of Education and Training grant number [B2023.DNA.03].

Institutional Review Board Statement

This research was funded and aggreement by the Ministry of Education and Training of Vietnam under grant number B2023.DNA.03. The authors would like to express their sincere gratitude for the support provided.

Informed Consent Statement

Not applicable.

Data Availability Statement

The author declares the data associated with a paper is available and open for collecting and processing. The author declares code associated with a paper is available.

Acknowledgments

This research is funded by Ministry of Education and Training under project number B2023.DNA.03.

Conflicts of Interest

The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

[1] Trần, T.D., 2012. Họ ngôn ngữ và văn hóa tiền sử: Trường hợp văn hóa Đông Sơn và họ Thái - Kađai. In: Cộng đồng các tộc người ngữ hệ Thái - Kađai ở Việt Nam. Nhà xuất bản Thế giới: Hanoi, Vietnam. pp. 337–346.

[2] Trần, T.D., 2009. Về mối quan hệ giữa các ngôn ngữ Nam Á và Nam Đảo ở Đông Nam Á. VNU Journal of Science: Social Sciences and Humanities. 25(3), 121–126.

[3] Po, D., 2006. Cham language and script in historical process. Proceedings of the History of Language and Cham Script; 21–22 September 2006; Kuala Lumpur, Malaysia. EFEO & Tokyo University of Foreign Studies: Paris, France; Tokyo Japan. pp. 1–10.

[4] Bùi, T.N., 2017. Another Hypothesis on the Semantics and Etymology of the Name 'Đà Nẵng'. Language & Life Journal. 9(263), 95–101.

[5] Trần, T.D., 2016. Stopization of initial sounds in the history of the vietnamese language. Journal of Linguistics. 5(324), 9–15.

[6] Trần, T.D., 2013. Tên gọi thánh "Dóng" và lễ hội "Phù Đổng": góc nhìn từ ngữ âm lịch sử tiếng Việt. Journal of Linguistics. 2(285), 3–10.

[7] Dinh, H.H., 2020. The symbol of Saint Gióng and the Gióng festival in the historical context of Vietnam. Asian Education and Development Studies. 9(1), 37–45. DOI: https://doi.org/10.1108/AEDS-01-2018-0015

[8] Parmentier, H., 1935. LA CONSTRUCTION DANS L'ARCHITECTURE KHMÈRE CLASSIQUE. Bulletin of the French School of the Far East (BEFEO). 35(2), 243-311.

[9] Hardy, A., Griffiths, A., Baptiste, P., et al., 2019. Champa Territories and Networks of a Southeast Asian Kingdom. Available from: https://publications.efeo.fr/en/livres/932_champa (cited 20 April 2025).

[10] Sastri, K.A., 1935. The Origin of the Champa Alphabet. Bulletin de l'École française d'Extrême-Orient. Bulletin of the French School of the Far East (BEFEO). 35, 233–241.

[11] Liu, C., Yin, F., Wang, D., et al., 2013. Online and offline handwritten Chinese character recognition: benchmarking on new databases. Pattern Recognition. 46(1), 155–162. DOI: https://doi.org/10.1016/j.patcog.2012.06.021

[12] Xiao, X., Jin, L., Yang, Y., et al., 2017. Building Fast and Compact Convolutional Neural Networks for Offline Handwritten Chinese Character Recognition. Available from: https://doi.org/10.48550/arXiv.1702.07975 (cited 20 April 2025).

[13] Wu, Y., & Yin, F., Chen, Z., et al., 2017. Handwritten Chinese Text Recognition Using Separable Multi-Dimensional Recurrent Neural Network. Available from: https://ieeexplore.ieee.org/document/8269953 (cited 20 April 2025).

[14] Zhang, Y., 2015. Deep Convolutional Network for Handwritten Chinese Character Recognition.

Available from: https://paperswithcode.com/paper/deep-convolutional-network-for-handwritten (cited 20 April 2025).

[15] Zhuang, Y., Liu, Q., Qiu, C., et al., 2021. A Handwritten Chinese Character Recognition based on Convolutional Neural Network and Median Filtering. Journal of Physics: Conference Series. DOI: https://doi.org/10.1088/1742-6596/1820/1/012162

[16] Huang, J., Tan, J., 2019. DropDist: A CNN-based model for large-scale handwritten Chinese character recognition with small training dataset. IEEE Transactions on Image Processing. 28(6), 2993–3003

[17] Ahlawat, S., Choudhary, A., 2020. Hybrid CNN-SVM Classifier for Handwritten Digit Recognition. Procedia Computer Science. 167, 2554-2560. DOI: https://doi.org/10.1016/j.procs.2020.03.309

[18] Kavitha, B.R., Srimathi, C., 2019. Benchmarking on offline Handwritten Tamil Character Recognition using convolutional neural networks. Journal of King Saud University - Computer and Information Sciences. 34(4), 1183-1190. DOI: https://doi.org/10.1016/j.jksuci.2019.06.004

[19] He, K., Zhang, X., Ren, S., et al., 2016. Deep Residual Learning for Image Recognition. Available from: https://doi.org/10.1109/CVPR.2016.90 (cited 20 April 2025).

[20] Liu, X., Zhou, Y., Wang, Z., 2020. Deep neural network-based recognition of entities in Chinese online medical inquiry texts. Future Generation Computer Systems. 114, 581-604. DOI: https://doi.org/10.1016/j.future.2020.08.022

[21] Peng, N., Dredze, M., 2016. Improving Named Entity Recognition for Chinese Social Media with Word Segmentation Representation Learning. Available from: https://doi.org/10.18653/v1/P16-2025 (cited 20 April 2025).

[22] Tuan, N.M., Meesad, P., 2025. A Bilinear Neural Network Method for Solving a Generalized Fractional (2+1)-Dimensional Konopelchenko-Dubrovsky-Kaup-Kupershmidt Equation. International Journal of Theoretical Physics. 64, 1–17. DOI: https://doi.org/10.1007/s10773-024-05855-w

[23] Tuan, N.M., Meesad, P., 2025. Bilinear Recurrent Neural Network for a Modified Benney-Luke Equation. International Journal of Applied and Computational Mathematics. 11(2), 1–35. DOI: https://doi.org/10.1007/s40819-025-01851-8

[24] Tuan, N.M., Meesad, P., Hieu, D.V., et al., 2024. On Students' Sentiment Prediction Based on Deep Learning: Applied Information Literacy. SN Computer Science. 5, 928. DOI: https://doi.org/10.1007/s42979-024-03281-7

[25] Tuan, N.M., Meesad, P., Nguyen, H.H.C., 2024. English–Vietnamese Machine Translation Using Deep Learning for Chatbot Applications. SN Computer Science. 5(1), 1–5. DOI: https://doi.org/10.1007/s42979-023-02339-2

[26] Papadakis, S., Kiv, A., Kravtsov, H., et al., 2023. Unlocking the power of synergy: The joint force of cloud technologies and augmented reality in education. Proceedings of the 10th Workshop on Cloud Technologies in Education, and 5th International Workshop on Augmented Reality in Education (CTE+AREdu 2022); 23 May 2022; Kryvyi Rih, Ukraine. 3364, pp. 1–23.

[27] Papadakis, S., Kiв, A.Ю., Kravtsov, H.M., et al., 2023. Revolutionizing education: Using computer simulation and cloud-based smart technology to facilitate successful open learning. Proceedings of the 10th Illia O. Teplytskyi Workshop on Computer Simulation in Education, and Workshop on Cloud-based Smart Technologies for Open Education (CoSinEi and CSTOE 2022) co-located with ACNS Conference on Cloud and Immersive Technologies in Education (CITEd 2022); 22 December 2022; Kyiv, Ukraine. 3358, pp. 1–18.

[28] Nguyen, M.T., Meesad, P., Nguyen, H.S., 2024. On a Stock Prediction Aligned to Natural Language Sentiments. Proceedings of the 2024 8th International Conference on Natural Language Processing and Information Retrieval; 13–15 December 2024; Okayama, Japan. ACM: New York, NY, USA. pp. 395–400.

[29] Nguyen, M.T., Phayung, M., Duong, V.H., et al., 2023. New data about library service quality and convolution prediction. CTU Journal of Innovation and Sustainable Development. 15(ISDS), 30–38. DOI: https://doi.org/10.22144/ctujoisd.2023.032