

Extensive Review of State-of-the-Art Classification Techniques for Cuneiform Symbol Imaging: Open Issues and Challenges

Farah Maath¹, Maha Mahmood¹, Belal Al-Khateeb¹^{*}

¹ College of Computer Science & Information Technology, Al Rumadi, Anbar, 31001, IRAQ

^{*}Corresponding Author: Belal Al-Khateeb

DOI: <https://doi.org/10.52866/ijcsm.2023.02.03.011>

Received May 2023; Accepted July 2023 ; Available online August 2023

ABSTRACT: The cuneiform script reveals some previously unknown aspects of our past. However, reading ancient clay tablets demands a substantial investment of time and persistent practice over a long period of time. As the fourth millennium came to a close, earlier recording methods gave way to the development of writing – the visual representation of spoken language. The first language to be transcribed in written form in Mesopotamia was Sumerian. Predominantly, the earliest tablets originate from the Uruk site in southern Mesopotamia, possibly marking its birthplace. Digitization cuneiform documents is imperative to boost research focused on the ancient Middle East. A few initiatives embarked upon this endeavor around the year 2000. Nonetheless, the digitization process is time-consuming due to the extensive volume of documents, and a dependable (semi) automatic methodology has yet to be developed. Given the antiquity of cuneiform script, recognizing cuneiform signs using real-world applications via two graph-based algorithms, each with complementary runtime characteristics, remains a manual procedure. Translating cuneiform proves to be a daunting task. Only in relatively recent times has grammar been established scientifically, while lexical challenges remain abundant and far from resolved. Furthermore, the majority of the Sumerian tablets have succumbed to the ravages of time, leaving behind only a handful of ancient depictions. Some of these old images have been preserved in a unique collection or in museums worldwide, allowing specialists to easily apply the sign detector to their cuneiform text studies. In this paper, we will discuss the categorization and analysis of clay tablets using a trained cuneiform model, employing artificial intelligence methodologies. Additionally, we will explore the methods employed, highlighting their strengths and weaknesses. Finally, we will propose potential directions for future research.

Keywords: Cuneiform, Sumerians Tablets, Artificial Neural Network, Deep Learning, Classification Algorithms

1. INTRODUCTION

At its inception, cuneiform signs were designed to represent sounds, enabling the recording of spoken language. This served to preserve language for posterity, allowing thoughts and concepts to be conveyed in written form. Among the earliest known forms of writing is cuneiform, etched onto clay tablets using a reed stylus with a wedge-shaped tip, hence its name [1]. Presently, 'cuneiform' denotes a writing system characterized by wedge-shaped markings. While numerous transliterated Akkadian texts that can be accessed for free in digitized corpora; however, there are a much smaller number of texts that have been phonologically transcribed. Employed for three millennia, cuneiform script emerged in the late fourth millennium BCE and persisted until the first century CE, producing a diverse array of texts by scribes. Though primarily etched onto clay, some found their place on wax-coated stone, metal, or wooden boards. Over a million cuneiform texts have been uncovered to date in a large Near Eastern region stretching from Anatolia to Iran and northern Iraq to Bahrain and Egypt [2]. This region includes a portion of the Arabian Peninsula and the Persian Gulf. The amalgamation of wedges gives rise to an array of cuneiform signs, given that the script is created by pressing a stylus into fresh clay. Cuneiform was developed in ancient Mesopotamia, the present methodology involves manual reconstruction of missing data [3]. Executed by a select group proficient in Akkadian and cuneiform, this labor-intensive process seeks to recover obscured lines in scientific, lexical, literary, and religious texts through comparison with intact tablets and parallel passages on deteriorated ones. Fortunately, the availability of duplicate copies, facilitates this endeavor. Nevertheless, it necessitates profound familiarity with various text types and collections, and the suggested restorations by scholars often invite diverse interpretations [4]. Though the system's brilliance permits effortless replication, scribes grappled with memorizing myriad signs. This discrepancy arises from the labor-intensive nature of transcription, which offers limited aid in academically interpreting cuneiform texts.

Numerous findings include literary works like the Epic of Gilgamesh, clay envelopes containing letters that were sealed with clay, and historical narratives. Remarkably, vast libraries, exemplified by the collection of Assyrian king Ashurbanipal (668-627 B.C.E.), have come to light [1]. An astronomical text dated to 75 CE is the most recent example

of cuneiform discovery. Over the course of its 3,000-year history, cuneiform has transcribed approximately 15 distinct languages, encompassing Akkadian, Babylonian, Sumerian, Hittite, Assyrian, Urartian, Elamite, and Old Persian [2]. Figure 1 illustrates ancient Sumerian cuneiform depictions. The main contribution of this paper lies in evaluating and showcasing the utility of artificial intelligence in translating cuneiform-inscribed clay tablets. This research is significant, forging connections between cuneiform studies and artificial intelligence, opening up new possibilities for deciphering and understanding ancient texts with enhanced efficiency and precision.



FIGURE 1. The 5,500-Year-Old Sumerian Cuneiform Tablets Reveal Stunning Revelations [1].

2. transliteration

Transliterating cuneiform necessitates adherence to a distinct pattern. This process entails choices made by the transliterating scholar, a result of the script's multiple meanings. Thus, for each sign, the scholar determines which among its several possible meanings aligns with the intended sense in the original document. For instance, the dingir sign within a Hittite text may signify the Hittite syllable, form an Akkadian phrase signifying the syllable, represent a Sumerian grammatical element denoting the determinative for a deity, or denote the original Sumerian sense of "god." The transliteration task selects a unique interpretation of the same glyph contingent on its role within the context [5]. Notably, Sumerian, Akkadian (Babylonian), and Hittite (as well as Luwian) cuneiform texts employ distinct transliteration systems. The convention of using acute and grave accents as abbreviations to disambiguate homophone is widely embraced across diverse disciplines. This practice sees extensive usage across the various fields. Hence, u corresponds to u1, the first glyph expressing the phonetic value of u. The glyph represented by an acute accent equates to the second, u2, while the grave-accented glyph, ù, corresponds to the third, u3 in the series (though the numerical sequence is traditional, it is fundamentally arbitrary and subject to historical decipherment). In Sumerian transliteration, typographic ligatures are indicated using the multiplication sign 'x.' The signs themselves are depicted in uppercase letters, while the specific chosen transliteration reading is depicted in lowercase letters. Therefore, uppercase letters are used to indicate "Diri" compound – a sequence of signs whose combined reading diverges from the sum of individual constituent sign readings. For instance, the compound IGI.A, uniting "eye" and "water" signs, has the reading imhur, signifying "foam" [6].

In a compound written in the Diri script, dots in the transliteration isolate individual signs. Uppercase letters can also signify a Sumerogram (for instance, K. BABBAR, meaning "silver" in Sumerian, employed with the intended Akkadian reading "kaspum," also meaning "silver"), an Akkadogram, or simply a sign sequence whose reading eludes the editor. The "real" reading, if deciphered, is naturally presented in italics within the transliteration. For example, IGI.A is rendered as imhur4. Contemporary estimates suggest the unearthing of between 500,000 and 2,000,000 cuneiform tablets, with a mere 30,000 to 100,000 having been read or published. Beyond the British Museum, notable collections include the Yale Babylonian Collection (about 40,000), the Louvre, Vorderasiatisches Museum Berlin, the National Museum of Iraq, the Istanbul Archaeology Museums, and the Penn Museum. The Yale Babylonian Collection stands as the world's second largest collection. Since there are only a few hundred qualified cuneiformists in the world, most of these texts have "lain in these collections for a century without being translated, studied, or published" [5].

3. Artificial Intelligent Classification Algorithms

Neural networks are one of the most popular classification methods, constituting a pivotal areas of research in data mining. Drawing inspiration from the human brain, Artificial Neural Networks (ANNs) serve as computational models. Recent strides in artificial intelligence encompass domains like voice recognition, image recognition, and robotics that make use of artificial neural networks. These networks consist of synthetic neurons inspired by biological counterparts and can be programmed to execute diverse functions. The process of training neural networks relies on the initial setting of certain parameters, including weight, bias, and the learning rate of the algorithm [7]. The optimization of

weight within an ANN is a combinatorial challenge, essential for attaining the desired output. The subsequent list outlines several learning methods for applicable to various classification predicaments within ANN:

a) ANN with Back Propagation (BP): One variant of ANN with BP is proposed in a neural network application for classification. This application makes use of neural networks. When training a neural network, the BP algorithm is typically utilized. In [8], an alternative BP-based ANN variant is presented, tailored for multispectral image classification. Following BP training on the classical portions of the image, the neural network is applied to the problem of image classification.

b) An enhanced version of the BP algorithm: This section explains training of a neural network through the BP algorithm by utilizing the gradient delta rule. It has a wide range of applications for the architecture of parallel hardware. Unlike conventional approaches, the momentum factor is recalculated at each step, fostering improved performance in both speed and convergence stability compared to traditional BP. Soft computing embraces various meta-heuristic algorithms, such as the Genetic Algorithm (GA), Particle Swarm Optimization (PSO), Firefly Algorithm (FA), and Cuckoo Search (CS) [9]. The training of neural networks can be accomplished with the help of these meta-heuristic algorithms, yielding approximations of the correct answers across diverse domains. Situations in which traditional algorithms produce a local optimum, these algorithms are used. A traditional algorithm not only raises the amount of money spent on computation but also takes more time to produce results. In previous research, numerous researchers have combined ANN with these meta-heuristic algorithms to circumvent the shortcomings of ANN. The examples include:

The Integration of PSO with ANN, known as PSNN, presents an evolutionary approach that combines architectural evolution with weight learning [10]. This innovative system addresses the performance enhancement of ANNs, particularly addressing limitations associated with structural methods like hill climbing, which often get stuck in local optima due to their dependency on initial network architecture. The PSNN hybrid method mitigates this issue and enhances results. In the realm of classification, a hybrid technique blending PSO and BP is proposed [9]. By leveraging PSO's global search capability and BP's local search prowess, this approach achieves higher classification accuracy compared to standalone BP. Simultaneously, it reduces CPU processing time. Application of this technique to iris pattern classification is highlighted. Fruit classification, as discussed in [11], introduces a hybrid methodology involving PSO, Artificial Bee Colony (ABC), and a single hidden layer feed Forward Neural Network (FNN).

c) ANN with GA: This section suggest an original structure for a hybrid neural network that can classify ECG beats [12]. It is used to determine the weight, including the number of nodes, in the first layer of the neural network. This study presents an application of GA in conjunction with neural networks for the purpose of land cover classification using remotely sensed data. It employs a real-coded GA hybrid that uses BP. Genetic operators prevent premature convergence, while initial connection weights are determined through the BP algorithm.

An ANN was used in conjunction with an ABC to propose a classification method for DNA microarrays. The dimensionality of a DNA microarray can be reduced using ABC, which is then used to select the best set of genes for identifying diseases. Finally, the ANN can be trained to classify the DNA microarray using these reduced genes. To determine whether an MR brain image is normal or abnormal, you should present a method that combines FNN and improved automatic binary classification (ABC). Improved ABC, which is founded on both fitness scaling and chaotic theory, is used to optimize the parameters of the Fuzzy Neural Network (FNN). The ABC algorithm was used to train a neural network to solve a classification problem relevant to the field of medicine, while the hybrid approach was used to solve nine different real-world problems related to the medical field [11].

d) ANN with Improved Cuckoo Search (ICS): This step involved the training of the neural network using ICS. The behavior of cuckoo species, in which they lay their eggs in the nests of host species, served as the inspiration for the cuckoo search. When it comes to the parameters, ICS is not the same as a standard cuckoo search. To find an improved solution both globally and locally, the parameters beta and alpha are utilized. It improves the algorithm's accuracy as well as the rate at which it converges. The default cuckoo search utilizes predetermined values for each of these parameters [13].

Training of ANN using Ant Colony Optimization (ACO): This step involved using ACO to optimize the weight of the neural network. The neural network is trained for pattern classification in the same way that the authors in [13], which trained the ANN using a hybrid technique that included both ACO and BP. BP that got caught up in a local optimal solution. Therefore, the purpose of this hybrid training is to employ a global optimization algorithm so that BP can begin with a healthy weight. Both are utilized for the classification of data in the medical field, specifically cancer datasets, diabetes datasets, and heart datasets.

e) ANN with Tabu Search: This system proposes the integration of four distinct neural network training methods: GA, simulated annealing, tabu search, and BP. The primary purpose of this system is to facilitate the training of neural networks. The uphill property is also present in simulated annealing (occasionally accept bad moves). On the other hand, the characteristic that distinguishes GA is parallel search, while tabu search is defined by its flexible memory [14]. Notably, all these characteristics are included in the system that has been proposed. This method is subsequently applied to address four classification problems and one prediction problem.

f) ANN with Gravitational Search Optimization (GSA): Presented here is a suggestion for an iris recognition system, employing the FNNPSO and FNNGSA hybrid classification methods for iris analysis. The procedural framework encompasses image acquisition, image segmentation, segmentation normalization, feature extraction, and feature-based classification using ANN. During the process of training neural networks to determine optimal weights and biases, both PSO and GSA are applied as training methods [15].

Using ANN combined with biogeography-based optimization, suggest a fruit classification method that makes use of the fruit's color, shape, and texture features. To keep the weight of the neural network up to date, a biogeography-based optimization algorithm is utilized [15].

g) ANN-Based Fuzzy Logic: This model proposes the utilization of neural networks in decision-making systems and fuzzy logic control. By training connectionist structures, fuzzy logic rules are established, and input-output relationships are identified [16].

4. CUNEIFORM SYMBOLS CLASSIFICATION

In this section, we will discuss works that have employed various artificial intelligence strategies to complete the classification of cuneiform symbols.

In 2003, Natalia Koslova and Peter Damerow [17] proposed that the transcription of ancient texts be based on a highly developed historical set of transliteration standards. This approach employs indices and diacritics to distinguish between homophones, as well as specialized signs and unique characters to indicate whether traces of cuneiform signs support the proposed transliteration reading. This system resisted all attempts to map it into electronic data types for an extended period. As the benefits of utilizing computer technology became increasingly evident, the formatting capabilities of text processors and typefaces with unusual characters began to dominate its use for expressing transliterations of cuneiform texts. This shift occurred due to the undeniable advantages of utilizing computer technology. These techniques were appropriate for adjusting the display and printing of electronic transliterations to traditional standards; however, they were of very little use for any sophisticated text processing that made use of language technologies. As this realization dawned, efforts were made to streamline the conventions system in such a way that pure ASCII text could accurately represent the transliterations. Despite these efforts, the situation became even more precarious. The traditional system's precision was compromised by the simplified ASCII transliterations to the extent that the latter could no longer be employed for traditional publications. In addition, the simpler ASCII transliterations did not promote the development and deployment of advanced language technology. The Cuneiform Digital Library Journal (CDLJ) serves as a refereed electronic journal for cuneiform studies under the umbrella of a non-profit organization. The Cuneiform Digital Library Initiative (CDLI) focuses on major themes such as text analysis of documents from the 4th and 3rd millennia, early language, writing, paleography, administrative history, mathematics, metrology, and modern cuneiform editing technology. These text analyses encompass diverse elements like text, photographs, data, drawings, and interpretations. However, they were not particularly useful for complex text processing utilizing language technologies.

In 2013, Naktal M. Edan [18] expressed an interest in developing an algorithm for the recognition of cuneiform symbols. First, the Sumerian texts are fed into the scanner to undergo preliminary preprocessing operations. These operations, such as segmentation, are performed with the intention of chopping up the text to obtain a cuneiform symbol. Subsequently, the features of each symbol are extracted by using the center of gravity, vertical and horizontal projections, and connected components. Given the multitude of cuneiform symbols, a K-means clustering algorithm is used to group symbols that share substantial similarities. Following this, multilayer neural networks are utilized to classify each symbol within its respective cluster. The optimal value, which yields a lower overall classification error rate, is chosen. Here, a total of five symbol classes are present. A separate multilayer neural network is constructed for each symbol class, with the number of inputs, hidden nodes, and output nodes in the network determined by the total number of symbols within the class under examination. During the testing phase, the initial step involves determining the classes of unknown cuneiform symbols by calculating the Euclidean distance between each centroid of the symbol's respective cluster and the symbol itself. Ultimately, the appropriate multilayer neural network is assigned the calculation results.

In 2015, Raed Majeed et al. [19] employed a novel method to obtain the cuneiform text from old Sumerian clay tablets. The method is based on automatic selection of wavelet bases. This approach offers cuneiformists and archaeologists a speedy, dynamic, and straightforward means of extracting cuneiform sentences. Using samples of images, the proposed system produces superior results. The enhancement algorithm presented is designed to address the primary issues that are present in cuneiform tablets. The process of deciphering the cuneiform writing on the Sumerian tablets required the development of an innovative method for the automatic selection of wavelet bases and parameters. The wavelet transforms pay attention to the features in both the frequency and space domain. This method effectively connected corroded symbols to extract additional cuneiform text, providing dependable technology to assist cuneiformists and

archaeological researchers. The proposed improvement algorithm is intended to address several flaws that are present in cuneiform tablets. For instance, the input cuneiform tablet image normalizes global statistics, which means that the contrast and brightness of the Sumerian tablet image are also being normalized because of this process. When decomposing an image with wavelet transforms, local information is utilized, and the maximum absolute value of each part of the coefficients is used to adjust the coefficients. The Corrosion method uses contextual filters' properties to link broken symbols in the text; this would result in more accurate translations in later stages of the process of language translation. The method that has been proposed has the potential to be very helpful for many applications involving image processing and pattern recognition. Additionally, it might eventually be used for Egyptian hieroglyphics as well as those written in Assyrian, Babylonian, and Akkadian. However, the wavelet algorithm suffers from shift sensitivity, poor directionality, and a lack of phase information because of this work's limitation, which is that there is no phase information.

In 2016, Md Jan and Nada A. Rasheed [20] proposed a method for reconstructing ancient objects based on 2D images, addressing the limitations of previous approaches. This method takes advantage of the characteristics of the fragments' edges as well as the colors that are located close to the fragments' edges. The primary objective of this approach is to determine the precise contours of the fragments in question and then treat those determinations as curves. Subsequently, a polynomial function calculates curve coefficients, efficiently recognizing other edges of corresponding fragments, a prerequisite before progressing to the next phase. Extracting color features along fragments edges prove particularly effective for matching fragments, given the increased color variation at edges. Coefficient correlation and color intersection combination identify corresponding edges of other fragments. Utilizing polynomial function coefficients and color intersections, this method achieves a high success rate. The proposed method comprises several procedures. The first step involves obtaining the 2D fragment images. Edge filtering is then performed in the subsequent stage. A polynomial function is applied to determine coefficient values. Finally, matching between similar edges pairs of different fragments occurs, based on coefficient values and color intersections. Implementing this approach necessitates utilizing a recognition technique to locate the edge pairs of various fragments requiring matching.

Based on numerous applications of the Support Vector Machine (SVM) classifier and its discriminative functions, achieving a high level of accuracy in the recognition process. Ali Adel Said et al. [21] demonstrated in 2018 that the polygon approximation method is optimal for feature extraction. This method was utilized for recognition task competition with the elliptic Fourier descriptor. Two distinct datasets were employed for this study. The first datasets include 320 images of cuneiform symbol patterns, and its purpose is to evaluate the most effective feature extraction method. The second dataset, which is used to test the recognition system, has 240 pictures of cuneiform characters, and is composed of 2D representations of four different types of triangular patterns. This study compares two feature extraction techniques, the polygon approximation and Elliptic Fourier Descriptor (EFD) methods, using SVM as a classifier model with its discriminative kernel functions. The EFD and polygon approximation methods have been used to extract features. The results of examined recognition indicated that the polygon approximation by dominant points is more accurate for use in classification models, as the accuracy achieved with the RBF kernel function is 94% compared to the EFD method. As a result, the analyzed recognition results indicate that the approximation of polygons by dominant points is more accurate for use in classification models. The recognition of the cunei-form letters rather than their translation was the focus of this study.

The researchers who created [22] came up with the idea for a computer-based information system project, enabling it to remain intact for millennia despite being buried underground. The Hittite language, present on cuneiform tablets from the Hittite era, was manually analyzed. Unfortunately, researchers in the field have only read and translated a relatively small number of the tablets that have been uncovered. Numerous tablets remain interred in Anatolia, awaiting interpretation and translation into various languages. It would be a significant contribution not only to Anatolian and Turkish history, but also to the history of humanity if the cuneiform signs could be read and translated with the assistance of computer-aided techniques. The authors report the recognition of Hittite cuneiform signs through the application of computer-based image-processing techniques. In addition to that, the paper contains examples of different data mining applications and their uses. The authors also demonstrated the viability of an expert system based on Hittite cuneiform scripts, which is perhaps the most important of their contributions. The reading of Hittite texts written in cuneiform scripts on clay tablets, their hand-copied versions, or photographs using computer-based sign recognition techniques is the primary goal of that project. Most importantly, the authors demonstrated that it would be possible to build an expert system based on the cuneiform scripts used by the Hittites. The benefits of the project that are anticipated to result from its completion are as follows: prompt response, high performance, understandability, and dependability. An area necessitating improvement in this endeavor pertains to the low resolution of the cuneiform sign images from Portal Mainz, posing challenges for the research being conducted. The limited resolution has hindered the visibility of certain sign details.

In 2018, Nada A. Rasheed and Wissam Lahmod Nados [23] presented the specifics of the proposed method aimed at recognizing cuneiform symbols using neural networks. The goal of this work is to develop a neural network-based method for cuneiform symbol recognition. To accomplish this goal, the image is initially loaded into the memory, followed by pre-processing can be performed, during which some modifications may be introduced. The attributes are then extracted by estimating the frequency of the on-pixels in the (50 by 50) image matrix along horizontal lines, vertical lines, upper triangle of the main diagonal, and lower triangle of the main diagonal. In conclusion, a recognition process using the BP algorithm is carried out in a relatively short amount of time, with 0.05 as the learning rate and 20 hidden nodes. Extracted features encompass four histograms (horizontal lines, vertical lines, upper triangle of the main diagonal, and lower triangle of the main diagonal), resulting in a favorable outcome. Application of the algorithm to a database containing standard Sumerian symbols yields 80% accuracy in distinguishing between Sumerian characters, signifying promising result. It is possible to draw the conclusion that this method can be utilized for the purpose of recognizing other objects since it is efficient and produces encouraging results. That indicates that perhaps the same idea as with the other patterns can be used and applied for recognition. It was discovered that a network with 21 hidden nodes yields good results, indicating that there is a need to adjust the number of hidden nodes to find the optimal number of hidden neurons for the network so that it can carry out its work to the best of its ability. The effectiveness of the proposed methodology to classify has been demonstrated using a neural network, in particular the BP network, which proved the efficacy in pattern recognition by comparing the obtained result to the results of various NN architectures. The elimination of noise might cause some components to disappear, which would mean that the classification would be incorrect. On the other hand, in most practical formulations for pattern recognition, a classification rule does not lead to perfection. There is always the possibility of making a mistake in classification.

In 2018, Kenji Yamauchi et al. [24] concentrated on the digitalization of cuneiform documents, necessitating the development of an Optical Character Recognition (OCR) method for handwritten copies of the original materials. Currently, with the assistance of experts, the construction of a handwritten cuneiform character image set is being done as the first step toward the development of such methods. This image collection depicts the normal stroke patterns for writing the most common cuneiform characters. It will aid in constructing an OCR system for handwritten cuneiform, extracting character candidates from tablet images, applying Gaussian smoothing, and detecting regions via enclosed contours using the same method that was used when extracting the tablet images. After removing candidates from consideration based on having an area that is less than a heuristically determined value, the remaining candidates are then resized to be 64 by 64. Due to these procedures, authors were able to derive 147,010 candidate images with a resolution of 64 by 64 pixels and grayscale. Subsequent to the image set construction, an attempt will be made to recognize cuneiform characters from a given handwritten source using image processing with Handwritten Cuneiform Character Collection (HCCC) and by taking linguistic context into consideration. The reason for this is to address the variety of languages written in cuneiform as well as the ambiguity of character glyphs. In particular, the latter problem needs to be addressed in both a graphic and a linguistic manner. Each cuneiform character has a phono grammatic equivalent as well as an ideographic equivalent, and many of the cuneiform characters are difficult to differentiate from one another.

The authors of [25] proposed two complementary graph-based methods for the recognition of cuneiform signs by presenting a graph model for cuneiform signs along with a tailored distance measure based on the concept of the graph edit distance, proposing efficient heuristics for its computation, and experimentally demonstrating its effectiveness in classification tasks. To accomplish this goal, a nearest neighbor classifier is implemented using the distance measure, resulting in an increased amount of computation required for the prediction phase as the size of the training set grows. To solve this problem, it has been suggested that CNNs should be modified so that they can be applied to graphs as a different approach. This would shift the computational cost to the training phase, and it would be necessary to demonstrate the practicability of both approaches through an experimental comparison of their respective runtimes and prediction accuracies. When the training set is enriched with augmented examples, CNNs can achieve a high level of accuracy even though the amount of annotated real-world data that is currently available is still relatively limited. However, this method's scalability to larger datasets is limited, and its feasibility may decrease as training data volume continues to rise due to the proportional increase in prediction time with training set size.

It was reported in 2019 by A. Ziyad Aktas et al. [26] that Hittite cuneiform signs could be recognized through the application of computer-based image-processing techniques. In addition, data mining applications have been incorporated into the paper. The authors also demonstrated the viability of an expert system based on Hittite cuneiform scripts, which is perhaps the most important of their contributions. The primary goal of that project is to read Hittite texts written in cuneiform scripts on clay tablets, hand-copied versions, or photographs using computer-based sign recognition techniques. The development of logical Hittite language grammar rules for loading into an expert system (or knowledge-based system), would eventually be used in the process of translating Hittite texts into Turkish, and then into English and German. 13 different computer-based sign recognition algorithms successfully used to read Hittite cuneiform signs during three consecutive M.S. studies completed over the last few years at Backend University-

Ankara, Turkey's Computer Engineering Department. The accuracy and speed of the various algorithms were compared in these studies. Following this step, some of the grammar rules for the Hittite language are studied, and then they are converted into "IF THEN" rules so that they can be loaded into an existing expert system shell. After that, the Hittite cuneiform writing system was utilized to translate the signs into Turkish words. In a study involving broken clay tablets, an ANN algorithm is also used to deal with the problem. The fact that the authors of this work did not include all the cuneiform rules in a tree structure for the data flow diagram is one of the work's limitations. The project will require additional time, additional human resources, and additional financial resources.

In 2020, Aleksy Sahala et al. [27] conducted research to investigate automated phonological transcription of transliterated corpora. As it is normalized according to the grammatical description of a given dialect and explicitly shows the Akkadian renderings for Sumerian logograms, the phonological transcription offers a form to represent Akkadian that is aesthetically pleasing from a linguistic perspective. Given that cuneiform text does not mark the inflection of logograms, the inflected form must be deduced from the context of the sentence. Experiments conducted utilizing a context-aware neural network model demonstrate unequivocal success in transcription for syllabic tokens, delivering a recall rate of 90% at 1 and 96% at 3. This is expected to significantly aid the semi-automatic phonological transcription process. Despite logograms yielding less impressive results compared to syllabic tokens, context modeling can still clearly outperform the baseline, which demonstrates that it is beneficial. The more complicated token-based context modelling produced very similar outcomes to the more straightforward character-based context modeling, aligning with the fact that indeclinable words like prepositions and other logograms tend to provide crucial clues for the inflectional information of nearby words. This is because prepositions do not change their meaning. However, this work has some limitations, including a relatively small dataset, a relatively high workload when preparing the document for processing, and some ambiguity and unwanted predictions that arise because of oddities in the training data. These limitations make this work less desirable.

In 2020, Tobias Dencker et al. [28] proposed a weakly supervised method. For the method to function, tablet images and their transliterations had to be lined up, allowing the detection of transliterated signs within the tablet image. To retrain the sign detector, researchers used these localized signs instead of annotations. In the context of cuneiform script, deep learning needs a large amount of training data in the form of bounding boxes around the signs. However, acquiring this type of data is not easily accessible and proves costly. The solution lies in utilizing existing transliterations, which represent the tablet content in a sign-by-sign Latin script format. Since these do not provide sign localization, authors propose the following approach as a less stringent form of supervision: align tablet images with their respective transliterations to locate the transliterated signs within the tablet image. To retrain the sign detector, authors then substitute these localized signs for annotations. The quality of the alignments is improved because of having a better sign detector. Combining these steps into an iterative process makes it possible to train a cuneiform sign detector using only transliterations of the text. The performance of the cuneiform sign detector, evaluated on a large collection of clay tablets from the Neo-Assyrian period, significantly improved by the addition of a small number of annotations, even though the method that was used functions poorly under supervision. Additionally, it is advisable to develop a web application for the analysis of clay tablets using a trained cuneiform sign detector. This initiative would enable specialists to directly apply the sign detector in their research on cuneiform texts. Utilizing current transliterations that are part of the Assyriologist procedure, the proposed method provides sign annotations for sign detector training. This eliminates the need for manual annotations, required by other methods. Iterative learning methods alternate between training a sign detector and generating sign annotations during each iteration. This is accomplished by finding a solution to the alignment problem between tablet image and transliteration across many clay tablets in parallel. The approach that focuses on improving the quality and quantity of the generated sign annotations is intended to make the sign detector more effective. The difficulties of transliteration alignment, the fine-grained nature of cuneiform sign identification, and the stark imbalance in sign code class frequency, the generation of sign annotations has been split into two distinct steps: sign placement and image-transliteration alignment. These steps were implemented to address these challenges. Strike a balance between learning from aligned detections that are dependable and learning from placed sign detections with sign code classes that are unpredictable or even unknown to the detector (exploration). In other words, learn from placed detections (exploitation). The exploration component of the approach is responsible for injecting new sign code classes and sign annotations into the training of the sign detector. This is achieved by methodically filling in the gaps left by the alignment stage with transliteration indications. The approach's exploitation component produces very accurate sign annotations that are consistent with both the information that is accessible from the top-down and the bottom-up (line and sign detections in tablet images) (transliteration and line-geometry prior). There is not enough training data available in the form of bounding box annotations, and collecting these annotations takes a lot of time and requires in-depth industry knowledge.

In 2020, Ethan Fetayaa et al. [29] conducted research focused on a method to automatically complete broken passages in Late Babylonian archival texts. This method harnesses contemporary machine learning techniques, specifically Recurrent Neural Networks (RNNs). A tokenization strategy for Akkadian transliterations has been designed, and its

specifics can be found in the Materials and Methods section. Training is performed on an LSTM recurrent network as well as an n-gram baseline model. Due to the small amount of digitized cuneiform texts now available, it is uncertain if such data-driven algorithms would produce credible restorations in all sorts of writings. These models should, however, function effectively for genres with highly structured syntax, such as the legal, economic, and administrative Late Babylonian writings, as exemplified here. While analogous unsupervised language modelling tasks can virtually amass an infinite number of texts online for English, the primary challenge lies in the computational difficulty involved in storing and processing extensive data volumes. For more information, refer to similar unsupervised language modeling tasks. The majority of the 44 corrections made by the algorithm primarily underscore its proficiency in recognizing correct sentence structure (category 1: questions 1 to 23, 25 to 27, 29 to 35, 38 to 42, 46, and 48 to 52). In essence, the algorithm adjusts syntactic sequences of parts of speech in accordance with the statistical frequency of smaller syntagmatic structures. In the answer ranking, a total of 34 restorations demonstrated accurate semantic identification of the noun class and related verbs (category 2: questions 1, 3, 6 to 13, 16 to 18, 22, 24 to 27, 29, 31 to 33, 35 to 40, 42, 46 to 49, and 51). In fact, identification of appropriate sentence structure is a feature that is shared by a sizeable portion of these cases—namely, 30 questions. The proposed model eliminates the word in the middle of the sentence by making use of a prediction method that makes use of the rest of the sentence; this results in an increase in the randomness of the outcomes.

In 2020, Cécile Michel [30] dealt with a wide range of examples of such cases, such as copies, replicas, imitations, transformations, and fakes, to gain an understanding of the context in which these examples are created, the motivations of their creators, and, whenever it was possible, how scholars and collectors approach these examples. Specifically, she focused on how these examples are handled. Although the system is brilliant and very simple to replicate, the scribes were required to memorize many different signs. Most collections worldwide contain examples of forged writing, either of historical or contemporary origin. Identifying antique fakes poses minimal challenge for Assyriologists, unlike modern fakes. The study addresses diverse scenarios where the originator's purpose is not always evident—such as creating a copy, replica, imitation, fake, or forgery. In the beginning, some historical examples are presented, and the discussion centers on determining whether these examples are fakes, as well as the motivations behind their creation. The second section discusses contemporary examples, ranging from the production methods to the detection procedures used to uncover them. Authors traced the history of their production beginning in the 19th century and continuing up to the present day to gain an understanding of how museums and academics have dealt with them. Modern fakes first appeared on the antique market at an early stage.

Arwa Hamed et al. presents a novel method in 2021 [31] for translating cuneiform writing by utilizing ANN. For all intents and purposes, a Multi-Layer Perceptron (MLP) neural network has been adapted to translate images of Sumerian cuneiform symbols into corresponding English letters. The process involves utilizing an artificial intelligence technique, the MLP neural network, to translate images of Sumerian cuneiform symbols into English letters. The central concept behind the method that authors have suggested is to acquire an image of any cuneiform symbol and then generate an indicator for the corresponding English letter. Considering this, it is possible for letters of the English alphabet to be intelligently generated from cuneiform symbol images. The proposed method uses neural networks, but it can also be represented using other information technology and artificial intelligence models. This work has been successfully established, and it achieved a score of 100%; however, it faced challenges such as the inability to recognize chipped circuit boards, lack of public access to the dataset, difficulty in determining network structure, and absence of pre-processing. Consequently, the proposed network is unable to deal with the variations that can be found in images, such as rotation, noise, and so on.

Table 1 summarizes the works that are done in cuneiform symbols classification. In the following section, we will discuss the datasets that were utilized in the research articles that were outlined earlier.

Table 1. The details, strengths, and weaknesses of the used classification algorithms.

Author(s)/ Year	Methods/Techniques	Brief Description	Dataset	Strengths	Weaknesses
Natalia Koslova et al. (2003) [17].	They used ASCII Transliteration Format ATF and XML document type description.	They have built a digital library (CDLI) which provides a working environment which facilitates scientific work on these data.	The Hemitage Museum Images.	The CDLI makes a few minutes of searching.	They were of little use for any sophisticated text processing using language technologies.
Raed Majeed et al. (2015) [19].	Bandpass filter. The wavelet algorithm (Haar, Daubechies, Coiflets, Symlets, and discrete Meyer wavelets).	They proposed a method based on the wavelet algorithm to obtain the text from the Sumerian clay tablet.	Cuneiform library at Cornell University (Garšana collections) From CDLI.	The proposed method is an easy, fast, and active method for extracting the cuneiform sentences.	Wavelet algorithm suffers shift sensitivity, poor directionality, and a lack of phase information.
Md Jan Nordin and Nada A. Rasheed (2016) [20].	Polynomial function (smooth curve). The correlation coefficient. The highest intersection of colors. Edge detection.	The paper proposes a method for reconstructing ancient objects based on the use of 2D images. The method exploits. The features of the edges of fragments.	Local dataset	Reconstruct archaeological objects accurately by using polynomial function	The proposed work needs the recognition technique for the identification of different fragments edges pairs to be matched. The dataset is limited and not available.
Nada A. Rasheed Wessam Lahmod Nados (2018) [23].	Image processing (noise removal), Backpropagation Neural Network (BNN) algorithm.	They proposed a method able to recognize cuneiform symbols through the use of BNN. the method has used image processing to produce the database of the image.	No dataset	Short training time.	Using Noise removal may be removed some parts thus the classification could be failed. However, in most practical pattern recognition formulation, a classification rule does not lead to perfection. There always a probability of classification error. The proposed work has not been tested on a general dataset. Low accuracy.
Yamauchi et al. (2018) [24].	This study used OCR method, using, and applied Gaussian smoothing and detecting regions by enclosed contours with the same method when they used for tablet images extracting.	Building a handwritten cuneiform character image set. Focused on a cuneiform document digitization task, realization of OCR method from the handwritten copies of original	HCCC.	Convenient system of optical recognition of most existing printed cuneiform character. Scalable by	Dataset has limited size. Expensive and time consuming.

		materials.		Expanding its features	
Nils M. Kriege et al. (2018) [25].	Euclidean distance, graph convolutional network Spline CNN, and Graph model.	Focus on the recognition of cuneiform script and propose a graph model embedding tailored distance measures while controlling the computation cost of the prediction phase, along with efficient heuristics for the demonstration and computation of classification tasks effectiveness.	Portal Mainz.	High classification accuracy 94% with the Spline CNN.	This approach does not readily scale to large datasets. Those approaches can be expected to become less feasible when the available training data grows.
A. Ziya Aktas et al. (2019) [26].	Image-processing techniques and data-mining classification algorithms, expert system (Exsys), and CORVID.	The applicability or the technical feasibility of using image processing and computer vision techniques and knowledge-based systems or expert systems on the translation of Hittite cuneiform scripts written on clay tablets or their copies are demonstrated. Data mining applications are also used.	Portal Mainz.	Grammar rules of Hittite language are converted into logical rules to be stored into expert system shells, and sign recognition is successfully performed	They don't include all rules of the cuneiform in a tree structure of the data flow diagram. More time, more human resources, and more money are needed. Accuracy is not mentioned in the paper.
Ali Adel Saaid et al. (2019) [21].	Image enhancement (ideal low pass filter), image segmentation, image labeling, EFD, polygon approximation, and SVM.	After image processing steps, SVM is employed for the cuneiform letters' recognition.	They use a virtual dataset cuneiforms character images from Iraqi museum.	Classification accuracy of 94%.	There was no translation of cuneiform letters.
Aleksi Sahala et al. (2020) [27].	The paper use LSTM neural networks with Adam optimizer.	They have investigated automated phonological transcription of the transliterated corpora to represent Akkadian. This paper attempts to document the transcribing of Akkadians using a context-aware neural network model automatically.	Open Richly Annotated Cuneiform Corpus (ORACC).	Transcribe syllabic tokens automatically at near human performance with 96% recall @ 3. It is related to the normalization of historical text where context can often be helpful.	There is a greater workload when preparing the document. Relatively small dataset. Oddities in the training data produced some ambiguity and unwanted predictions.

Ethan Fetaya et al. (2020) [29].	They used RNN and LSTM.	They were investigating the possibility of assisting scholars and even automatically completing the breaks in ancient Akkadian texts from Achaemenes period Babylonia by modelling the language using RNN.	Achemenet website.	Good in teasing out sentence structures by using trained LSTM models to complete missing words.	The proposed model uses a prediction method to remove the middle word that predicts using the rest of the sentence, thus increasing the randomness of the results. Accuracy can be further enhanced.
Tobias Dencker et al. (2020) [28].	Character based approach, Sign-level bounding boxes, Line detection method, and CNN.	They proposed a deep learning-based sign detector that locates and classifies cuneiform signs in images of clay tablets.	State Archives of Assyria online (SAAO) and CDLI.	Full iterative training of a cuneiform sign detector. Increase precision by aligning TP detections while discarding FPs by using the weak supervision from transliterations.	The collection of the sufficient amount of training data is not available as will required more time, and expert knowledge.
Arwa Hamed et al. (2021) [39].	ANN by MLP neural network.	This study presents a new approach to translate the cuneiform writing by employing ANN. MLP neural network has been adapted for translating the Sumerian cuneiform symbol images to their corresponding English letters	Abd Al-Rahman Abd Al-Latif Al-Namir dataset.	Error value of 0.0009999.	It was not able to identify chipped boards. The dataset is not publicly available. There was a difficulty in determining the structure of the network. There is no preprocessing done, this means that the proposed network can't handle the variations that can be found in images like rotation, noise, and etc.

5. DATASETS

A dataset encompasses a collection of interrelated observations that are systematically organized and formatted to serve a specific purpose. Beyond the conventional tabular format, a dataset can also encompass documents, graphs, or images. Therefore, dataset search includes the process of discovering datasets, exploring their contents, and returning them to end-user [32]. Nevertheless, the focus of this paper is on the cuneiform symbol (characters, text, corpus, signs, and scripts). In the following subsections, insights into several datasets of cuneiform symbols used in the classification

studies of detecting cuneiform characters of the existing hand copies of the tablets are provided. Access to all the datasets listed in Table 2 can be gained by clicking on the URL links that are provided for each set in the table.

Table 2. Datasets for cuneiform symbol classification.

Dataset	No. of samples	No. of classes	Used by	Reference
CDLI	500,000	Three	[19], [28], [32]	https://cdli.ucla.edu/
HCCC	147,010	Fifteen	[24]	https://github.com/project-abgal/hccc
ORACC	21,078	Seven	[27]	http://oracc.museum.upenn.edu/
SAAo	5056	Two	[28]	http://oracc.museum.upenn.edu/saa/
HPM	500,000	Thirty	[25],[26]	https://www.hethport.uni-wuerzburg.de/HPM/index.php?lang=EN
Achaemenes	220,926	-	[29]	http://www.achemenet.com/en/

A. THE CUNEIFORM DIGITAL LIBRARY INITIATIVE (CDLI)

The Cuneiform Digital Library Initiative (CDLI) is actively constructing an expanding network of data, encompassing representations of cuneiform tablets, including their structures, contents, and results of academic research. The working environment that the library provides makes it easier for scientists to do research based on these data. The primary objective of CDLI is to digitally preserve and offer internet accessibility to all cuneiform inscriptions spanning from approximately 3350 B.C. before the inception of writing, to the conclusion of the pre-Christian era. These inscriptions emerged between the genesis of writing and the culmination of the pre-Christian era. The number of these artifacts currently housed in public and private collections is estimated to exceed 500,000 exemplars, of which the CDLI has now cataloged more than 320,000 in electronic format. CDLI represents the efforts of an international group of Assyriologists, museum curators, and historians. Online database maintained by the CDLI contains catalogue information for approximately 335,600 cuneiform text artifacts (along with 8,800 entries representing composite texts), 220,000 images, and 2.93 million lines of transliteration and related annotation [32]. A few examples of images taken from the CDLI dataset are displayed in Figure 2.

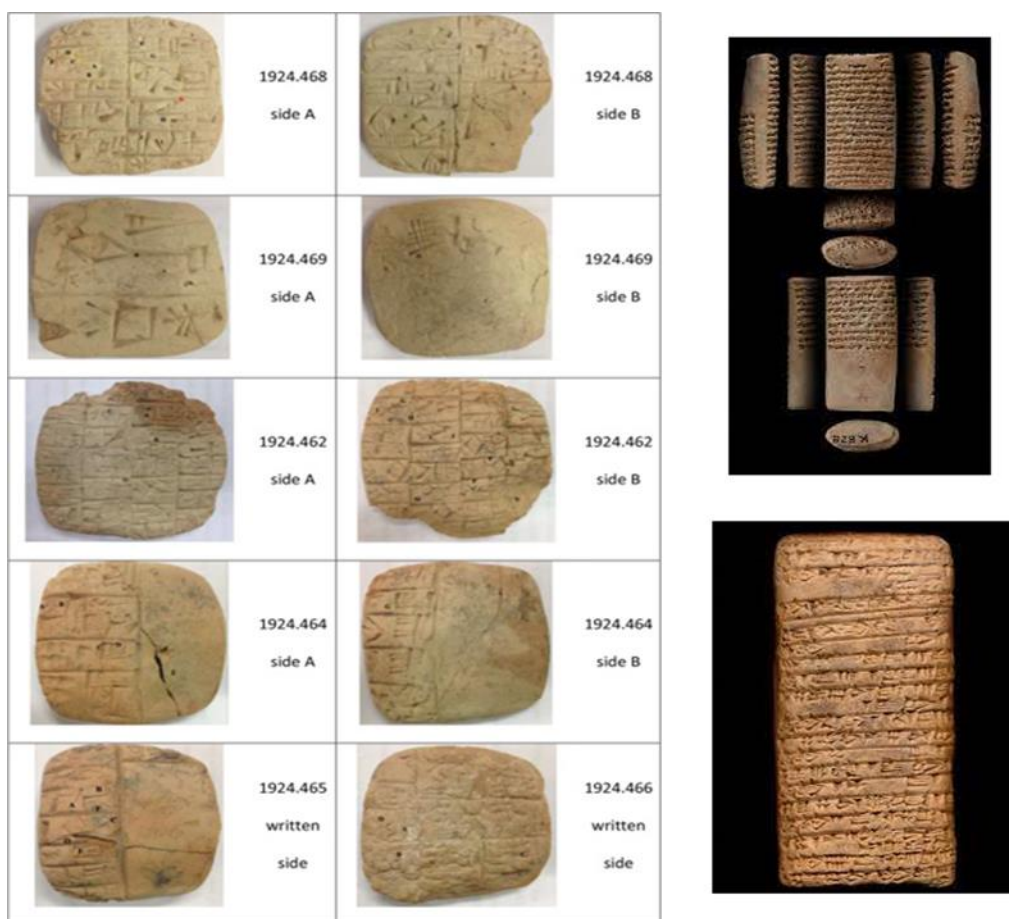


FIGURE 2. Samples of CDLI dataset [32].

B. HANDWRITTEN CUNEIFORM CHARACTER COLLECTION (HCCC)

HCCC stands as a specialized image compilation meticulously crafted for the specific purpose of developing an OCR system adept at discerning cuneiform characters from existing hand copies of tablets. The dataset, currently under construction, covers a comprehensive spectrum of 50 distinct classes of cuneiform characters. The selection criteria employed for assembling the HCCC dataset revolved around identifying the most recurrent classes across multiple corpora, thus ensuring a representative collection of commonly encountered cuneiform characters. Furthermore, there is a specific emphasis on the Ur III type (Neo-Sumerian) glyphs, as they play a significant role in cuneiform studies. The HCCC dataset comprises candidate images of cuneiform characters, totaling 147,010 in number. Each character class organized within its designated directory, denoted by labels such as A and AN, where the corresponding image files can be found. The images are formatted as 64 by 64 grayscale PNG files. It is important to note that the images in the HCCC dataset have undergone manual cleaning by experts in the field, including Assyriologists and Hittitologists, ensuring their accuracy and quality. The manual arrangement of the images further enhances the dataset's usability. For a more detailed breakdown of the current number of images collected for each character class, please refer to Figure 3 [24]. This breakdown provides valuable statistics regarding the composition and distribution of the HCCC dataset. By providing these additional details, you offer clearer descriptions of the HCCC dataset, including the number of classes (50) and the selection criteria based on frequency and the emphasis on Ur III type glyphs.

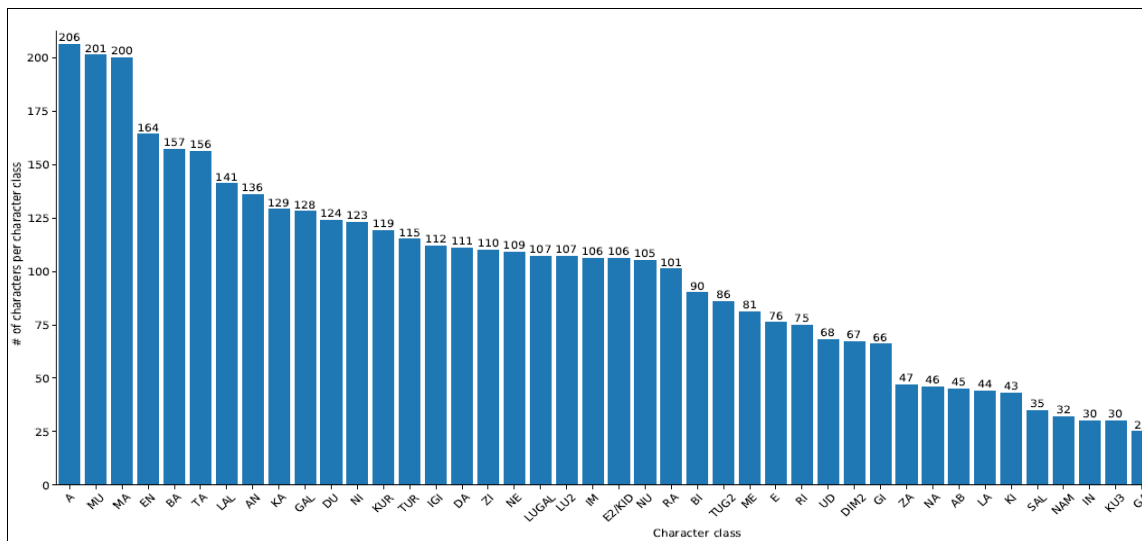


FIGURE 3. Distribution of HCCC collected image set [24].

C. OPEN RICHLY ANNOTATED CUNEIFORM CORPUS (ORACC)

ORACC serves as a natural complement to CDLI's pivotal archival role, functioning as an extensive catalog and image database of the entire cuneiform corpus, along with a repository for reference transliterations. ORACC was developed as a seamless expansion of CDLI, ORACC systematically receives automated transliterations from the CDLI database. The residing within ORACC are associated with tablets and/or composite texts, and their metadata provides valuable information about the genre of the text, the era in which it was written, and its provenance (the region in which the text was discovered). These details are essential considerations for researchers studying cuneiform texts. The fundamental building blocks of ORACC XML files are transliterations of words. These transliterations, written in Latin script, serve as representations of cuneiform signs. Even when no supplementary word-related details are provided, these transliterations are included in the files. In the context of cuneiform symbol classification research, a training set, development set, and test set were randomly selected from the ORACC dataset. This dataset encompasses a total of 21,078 sentences and contains 337,260 tokens. The training set consists of 271,311 tokens, the development set consists of 31,884 tokens, and the test set consists of 34,065 tokens [27]. To collect the texts included in the ORACC corpus, a list of ORACC projects was used. You can refer to Figure 4 [33] for a display of the ORACC projects that were utilized in gathering the texts. By incorporating these additional words and phrases, the revised text includes the relevant details about ORACC, its relationship with CDLI, metadata available for the texts, fundamental building blocks of ORACC XML files, and selection of the training, development, and test sets.

Project (abbreviation used in Oracc)	SUX	OLB	MPB	STB	NEB	LTB	NEA
Bilinguals in Late Mesopotamian Scholarship (blms)	x	x		x			
CAMS/Anzu (cams-anzu)				x			
CAMS/Barutu (cams-barutu)		x		x			
CAMS/The Standard Babylonian Epic of Etana (cams-etana)		x					x
CAMS/Geography of Knowledge Corpus (cams-gkab)	x			x		x	x
CAMS/Ludlul (cams-ludlul)				x			
CAMS/Seleucid Building Inscriptions (cams-selbi)				x			
Cuneiform Commentaries Project on ORACC (ccpo)	x			x			
Corpus of Kassite Sumerian Texts (ckst)	x				x		
The Amarna Texts (contrib-amarna)			x	x			
Cuneiform Texts Mentioning Israelites, Judeans ... (ctij)					x	x	x
Lexical Texts in the Royal Libraries at Nineveh (dcclt-nineveh)	x			x			
Reading the Signs (dcclt-signlists)	x			x			
Digital Corpus of Cuneiform Lexical Texts (dclt)	x	x	x	x			
Digital Corpus of Cuneiform Mathematical Texts (dcmt)	x	x		x			
Electronic Text Corpus of Sumerian Royal Inscriptions (etsri)	x						
Corpus of Glass Technological Texts (glass)				x			
Hellenistic Babylonian: Texts, Iconography, Names (hbtin)							x
Law and Order: Cuneiform Online Sustainable Tool (lacost)	x						
Old Babylonian Model Contracts (obmc)	x						
Old Babylonian Tabular Accounts (obta)	x	x					
The Inscr. of the Second Dynasty of Isin (ribo-babylon2)	x						
The Inscr. of the Period of the Uncertain Dynasties (ribo-babylon6)	x						
Rim-Anum: The House of Prisoners (rimanum)	x	x					
The Correspondence of Sargon II, Part I (saa0-saa01)							x
Neo-Assyrian Treaties and Loyalty Oaths (saa0-saa02)				x	x		x
Court Poetry and Literary Miscellanea (saa0-saa03)					x		x
Queries to the Sun God (saa0-saa04)					x		
The Correspondence of Sargon II, Part II (saa0-saa05)							x
Legal Trns. of the Royal Court of Nineveh, Part I (saa0-saa06)							x
Imperial Administrative Records, Part I (saa0-saa07)							x
Astrological Reports to Assyrian Kings (saa0-saa08)				x	x		x
Assyrian Prophecies (saa0-saa09)							x
Letters from Assyrian and Babylonian Scholars (saa0-saa10)				x	x		x
Imperial Administrative Records, Part II (saa0-saa11)							x
Grants, Decrees and Gifts of the Neo-Assyrian Period (saa0-saa12)							x
Letters from Assyrian and Babylonian Priests to ... (saa0-saa13)					x		x
Legal Trns. of the Royal Court of Nineveh, Part II (saa0-saa14)							x
The Correspondence of Sargon II, Part III (saa0-saa15)							x
The Political Correspondence of Esarhaddon (saa0-saa16)							x
The Neo-Babylonian Correspondence of Sargon and ... (saa0-saa17)					x		
The Babylonian Correspondence of Esarhaddon and ... (saa0-saa18)					x		
The Correspondence of Tiglath-Pileser III and ... (saa0-saa19)					x		x

FIGURE 4. The list of ORACC projects [33].

D. STATE ARCHIVES OF ASSYRIA ONLINE (SAAO)

SAAO is a web-based resource offering open access with the mission of making the rich Neo-Assyrian materials found in the royal archives of Nineveh and other locations more accessible to a wider audience. Since August 2015, SAAO has been a part of the Munich Open-access Cuneiform Corpus Initiative (MOCCI; directed by Karen Radner and Jamie Novotny) based at LMU in Munich. The mission of SAAO is to make the 5056 Neo-Assyrian texts published in the State Archives of Assyria Series available online for free in a format that is both fully searchable and richly annotated. In addition to that, to actively promote, facilitate, and broaden academic and other uses of these significant cuneiform sources, as well as to widely disseminate them. In addition, in their corpus they only included tablets that had a transliteration and an image available for use. Clay tablets would typically have writing on both sides, and sometimes even on the edges [34].

The information in the dataset is arranged according to collections of clay tablets. The organization of SAAO is followed for the most part, with the exception of the train and test collections, which contain some deviations. They offer a total of ten collections, eight of which are from SAAO and two custom collections that contain clay tablets from a variety of sources. These collections are as follows: saa01, saa05, saa06, saa08, saa09, saa10, saa13, saa16; train, and test. This dataset is kept in CSV files, and each of those files contains all of the information that pertains to one of the collections. The electronic text editions of SAAO and MOCCI are implemented within ORACC, which also hosts these editions. Figure 5 shows the process of translating a cuneiform text in one of the SAAO projects.

SAAO General structure:

- Data: tablet images, annotations, transliterations, and metadata.
- Experiments: training, testing, evaluation, and visualization.
- Lib: project library code.
- Results: generated detections (placed, raw, and aligned), network weights, and logs.
- Scripts: scripts to run the alignment and placement step of iterative training.

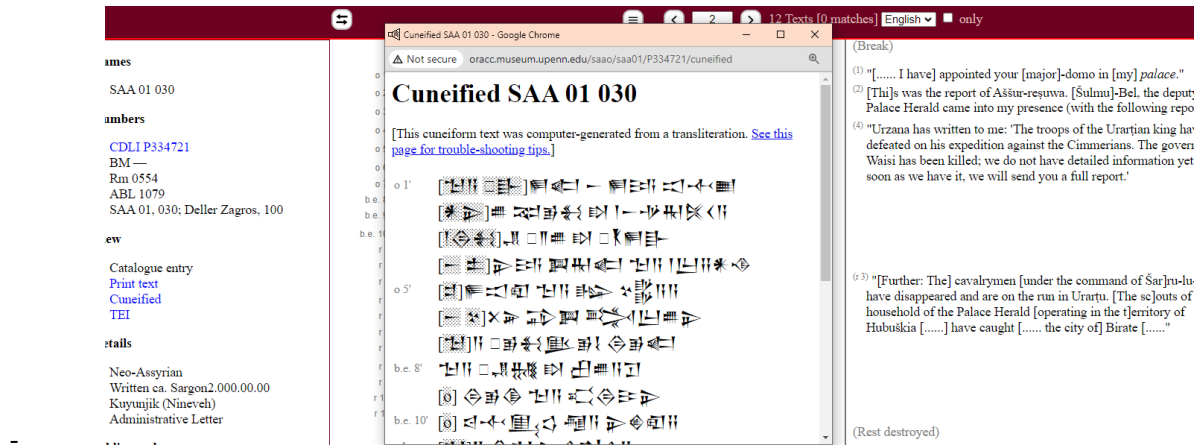


FIGURE 5. The process of translating a cuneiform text in one of the SAAO project [28].

E. HITTLE PORTAL MAINZ WEBSITE (HPM)

This database provides an online accessible database, housing an extensive catalog of recognized cuneiform tablets. This repository encompasses photographic collection, join-sketches, transcriptions, transliterations, publication data, and provenience information [35]. The portal also provides access to images of Hittite texts, with a goal of approximately 50,000 images (there are future plans for 3D rotational views of tablets as well as a Hittite bibliography). The Mainz Portal offers links or access to the digitalization projects that are being undertaken for Hittite textual and iconographic sources [36]. The HPM has a total of 5,680 vertices and 23,922 edges, and it is composed of 267 cuneiform signs. Figure 6 depicts the characteristics of the dataset, whereas Figure 7 displays a copy of a Hittite cuneiform script tablet that was used during [36].

Class	‘ba’	‘bi’	‘bu’	‘da’	‘di’	‘du’	‘ha’	‘hi’	‘hu’	‘ka’	‘ki’	‘ku’	‘la’	‘li’	‘lu’
$\#G$	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9
$ V $	36	24	16	8	16	8	20	16	16	20	16	28	24	28	16
$ E $	180	102	60	26	60	26	80	60	60	80	60	126	102	126	60
$h_{\max}$	7.5	7.4	10.0	5.3	7.5	7.7	7.3	7.9	7.5	9.1	9.1	7.4	7.7	6.7	7.5
$w_{\max}$	12.7	11.7	13.4	7.3	14.7	8.7	13.9	10.0	9.4	11.1	5.8	13.7	11.4	10.6	15.0
Class	‘na’	‘ni’	‘nu’	‘ra’	‘ri’	‘ru’	‘sa’	‘si’	‘su’	‘ta’	‘ti’	‘tu’	‘za’	‘zi’	‘zu’
$\#G$	9	9	9	9	9	9	9	9	9	9	9	9	8	8	8
$ V $	16	24	16	20	20	20	20	24	20	16	28	28	24	28	36
$ E $	60	102	60	80	80	80	80	102	80	60	126	126	102	126	180
$h_{\max}$	8.5	9.9	10.2	7.7	8.9	8.2	10.0	9.2	8.2	8.3	12.6	10.5	9.1	8.9	10.3
$w_{\max}$	14.4	7.6	8.6	9.6	11.5	10.2	10.4	11.2	11.7	8.4	15.2	12.4	9.2	13.2	21.0

FIGURE 6. HPM Dataset characteristics [36].

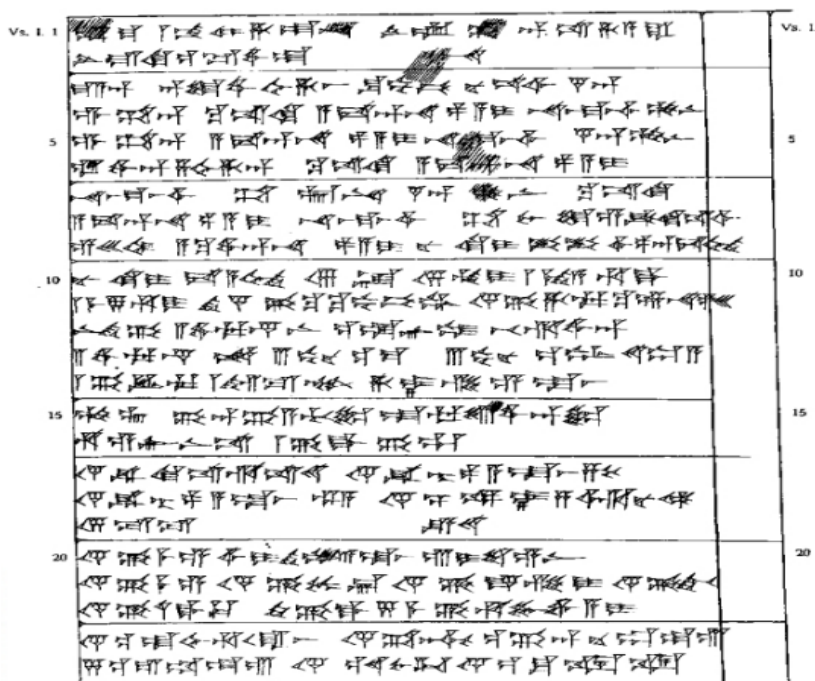


FIGURE 7. Hittite cuneiform script tablet copy [36].

F. ACHEMENET WEBSITE

Achemenet stands as a notable French online platform dedicated to the exploration of Persian history, archaeology, and Mesopotamian culture. This digital repository is centered around a comprehensive electronic database encompassing an array of materials written in Akkadian, specifically the Late Babylonian dialect. This dialect was used from the end of the cuneiform era, roughly in the first century CE, until the founding of the Neo-Babylonian empire in 627 BCE. All images and records shared with the Louvre are being added to the Oriental Institute's Integrated Database as part of its own online collections initiative, allowing for online searches [37]. Following this, 1,549 new terms were added to the vocabulary, bringing the total number of words in the collection to 220,926. 3,175 words appear just once, and 932 words appear twice [38]. Here are some key aspects of cuneiform covered on the website:

1. **Cuneiform Tablets:** The Achemenet website features a collection of cuneiform tablets, showcasing their diversity in terms of content, language, and historical period. These tablets include administrative records, legal documents, literary texts, and correspondence, offering insights into various aspects of ancient life.
2. **Decipherment and Transliteration:** The website discusses the decipherment of cuneiform, highlighting the efforts of scholars to decode the script and understand the languages written in cuneiform. It explains the transliteration process, which involves representing cuneiform signs using Latin characters.
3. **Cuneiform Languages:** Achemenet provides information on the different languages written in cuneiform, such as Sumerian, Akkadian, Elamite, and Old Persian. It explores their linguistic features, historical context, and textual sources.
4. **Cuneiform Literature:** The website delves into cuneiform literature, including myths, epic tales, hymns, and royal inscriptions. It offers translations and analyses of significant cuneiform literary texts, shedding light on the cultural and religious beliefs of the ancient Mesopotamians.
5. **Cuneiform and Archaeology:** Achemenet explores the connection between cuneiform and archaeology, discussing how cuneiform texts have contributed to our understanding of ancient societies. It explains how archaeological excavations have unearthed cuneiform tablets, providing valuable historical and cultural information.
6. **Cuneiform Resources:** The Achemenet website provides a range of resources for further exploration of cuneiform. These include bibliographies, dictionaries, and online databases that offer access to cuneiform texts and research materials.

Overall, the Achemenet website offers a comprehensive and in-depth exploration of cuneiform, providing a wealth of information on its history, decipherment, languages, literature, and archaeological significance.

6. CONCLUSIONS

The Sumerian language stands as a unique and isolated linguistic form, seemingly devoid of connections to other living or extinct languages. Cuneiform is a system of writing, identified by the wedge-shaped marks it makes on clay tablets. Cuneiform translation is a very difficult process, even for those who specialize in archaeological and cuneiform studies. The fragile nature of dry clay material makes it exceptionally susceptible to moisture and corrosion, leading to wear, damage, or the complete erosion of some inscriptions. On the other hand, whether they are modern or traditional, the methods that are used to archive the cuneiform, have some drawbacks in terms of reliability, accuracy, accessibility, time, usability, and the cost of the necessary hardware. There are only a few archives located across various digital libraries, special collections, and museums. Additionally, in 2004, the Baghdad Museum made the recent discovery of over 200,000 previously unidentified tablets. Most of the time, the quality of the Sumerian tablets that were provided was subpar.

In this paper, a review of the research that has been done in the field of cuneiform symbol classification was conducted, with a particular emphasis on the dataset that was used, classification approaches, limitations, and potential future research directions. This review concludes that there are potential research directions that could be taken in this area, primarily focusing on attempting to provide datasets and improving the accuracy of classification.

Funding

"This research received no external funding."

ACKNOWLEDGEMENT

None

CONFLICTS OF INTEREST:

"The authors declare no conflict of interest."

REFERENCES

- [1] E. Uchida and R. Watanabe et al., "Blackening of the Surfaces of Mesopotamian Clay Tablets due to Manganese Precipitation," *Archaeological Discovery*, vol. 2, no. 04, pp. 107, 2014, Scientific Research Publishing.
- [2] I. L. Finkel, "A join to the Map of the World: a notable discovery," *British Museum Magazine: The Journal of the British Museum Friends*, vol. 23, pp. 26-27, 1995.
- [3] W. Horowitz, "Mesopotamian Cosmic Geography," *Eisenbrauns*, vol. 8, 1998, Eisenbrauns.
- [4] R. M. Cormack, "A Review of Classification," *Journal of the Royal Statistical Society: Series A (General)*, vol. 134, no. 3, pp. 321-353, 1971. [Online]. Available: Wiley Online Library.
- [5] G. Yushu, "The Sumerian Account of the Invention of Writing. A New Interpretation," *Procedia-Social Behav. Sci.*, vol. 2, no. 5, pp. 7446–7453, 2010.
- [6] F. N. H. Al-Rawi and A. R. George, "Tablets from the Sippar Library, II. Tablet II of the Babylonian Creation Epic," *Iraq*, vol. 52, pp. 149–157, 1990.
- [7] M. Mahmood and B. Al-Khateeb, "Towards an Automatic Generation of Neural Networks," *J. Theor. & Appl. Inf. Technol.*, vol. 95, no. 23, 2017.
- [8] M. Mahmood, B. Al-Khateeb, and W. M. Alwash, "A Review on Neural Networks Approach on Classifying Cancers," *IAES Int. J. Artif. Intell.*, vol. 9, no. 2, p. 317, 2020.
- [9] W. J. AL-kubaisy, M. Yousif, B. Al-Khateeb, M. Mahmood, and D.-N. Le, "The Red Colobuses Monkey: A New Nature-Inspired Metaheuristic Optimization Algorithm," *Int. J. Comput. Intell. Syst.*, vol. 14, no. 1, pp. 1108–1118, 2021.
- [10] M. Mahmood and B. Al-Khateeb, "The Bluemonkey: A New Nature Inspired Metaheuristic Optimization Algorithm," *Period. Eng. Nat. Sci.*, vol. 7, no. 3, pp. 1054–1066, 2019.

- [11] B. Al-Khateeb, K. Ahmed, M. Mahmood, and D.-N. Le, "Rock Hyraxes Swarm Optimization: A New Nature-Inspired Metaheuristic Optimization Algorithm," *Comput. Mater. Contin.*, vol. 68, no. 1, pp. 643–654, 2021.
- [12] H. Mara, S. Krömker, S. Jakob, and B. Breuckmann, "GigaMesh and Gilgamesh: 3D Multiscale Integral Invariant Cuneiform Character Extraction," in *Proceedings of the 11th International Conference on Virtual Reality, Archaeology and Cultural Heritage*, pp. 131–138, 2010.
- [13] W. J. AL-kubaisy and M. Mahmood, "Classification of Texture Using Random Box Counting and Binarization Methods," *Bull. Electr. Eng. Informatics*, vol. 10, no. 1, pp. 533–540, 2021.
- [14] A. Hertz, "Tabu Search for Large Scale Timetabling Problems," *Eur. J. Oper. Res.*, vol. 54, no. 1, pp. 39–47, 1991.
- [15] B. Bogacz and H. Mara, "Automatable Annotations-Image Processing and Machine Learning for Script in 3D and 2D with Gigamesh," *Kodikologie und Paläographie im Digit. Zeitalter*, pp. 137–149, 2017.
- [16] Y. Kim, Y. Jernite, D. Sontag, and A. M. Rush, "Character-Aware Neural Language Models," in *13th AAAI Conference on Artificial Intelligence*, 2016.
- [17] N. Koslova and P. Damerow, "From Cuneiform Archives to Digital Libraries: The Hermitage Museum Joins the Cuneiform Digital Library Initiative," in *Proceedings of the 5th Russian Conference on Digital Libraries RCDL* pp. 1–8, 2003.
- [18] N. M. Edan, "Cuneiform Symbols Recognition Based on K-Means and Neural Network," *AL-Rafidain Journal of Computer Sciences and Mathematics*, vol. 10, no. 1, pp. 195-202, 2013. [Online]. Available: Mosul University.
- [19] R. Majeed, B. Z. Beiji, H. Hiyam, and W. Jumana, "Ancient Cuneiform Text Extraction Based on Automatic Wavelet Selection," *International Journal of Multimedia and Ubiquitous Engineering*, vol. 10, no. 6, pp. 253-264, 2015.
- [20] J. Nordin and N. A. Rasheed, "Reconstruction of Ancient Two-Dimensional Objects," *Int'l Journal of Computing, Communications & Instrumentation Engg. (IJCCIE)*, vol. 3, no. 2, pp. 1-5, June 2016, ISSN 2349-1469, EISSN 2349-1477, <http://dx.doi.org/10.15242/IJCCIE.U0516501>.
- [21] A. M. S. Rahma, "Cuneiform Symbols Recognition by Support Vector Machine (SVM) University of Technology / Department of Computer Sciences," January 2018, 2019.
- [22] A. Z. Aktas and T. Asuroglu, "Computerized Hittite Cuneiform Sign Recognition and Data Mining Application Examples," *Eur. Int. J. Sci. Technol.*, 2017.
- [23] N. A. Rasheed and W. L. Nadoos, "Recognition of Cuneiform Symbols Using Neural Network," *Journal of Theoretical and Applied Information Technology* , 15th September Vol.96. No 17, 2018.
- [24] K. Yamauchi, H. Yamamoto, and W. Mori, "Building A Handwritten Cuneiform Character Imageset," In *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC 2018)*. pp. 719–722, 2018.
- [25] N. M. Kriege, "Recognizing Cuneiform Signs Using Graph Based Methods," In *International Workshop on Cost-Sensitive Learning* (pp. 31-44). PMLR.vol. 1, pp. 31–44, 2018.
- [26] A. Z. Aktas and T. Asuroglu, "Computerized Hittite Cuneiform Sign Recognition and Knowledge-Based System Application Examples," vol. 15, no. 33, ,pp. 32–54, 2019.
- [27] A. Sahala, M. Silfverberg, A. Arppe, and K. Lind, "Automated Phonological Transcription of Akkadian Cuneiform Text Words ed," In *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*. European Language Resources Association (ELRA), No. May, pp. 3528–3534, 2020.
- [28] T. D. Id, P. Klinkisch, and S. M. Maul, "Deep Learning of Cuneiform Sign Detection with Weak Supervision Using Transliteration Alignment," *Plos one.*, vol. 12 no. 15 pp. 1–21, 2020.
- [29] E. Fetaya, Y. Lifshitz, E. Aaron, and S. Gordin, "Restoration of Fragmentary Babylonian Texts Using Recurrent Neural Networks," *Proceedings of the National Academy of Sciences*, 117(37),pp. 22743-22751 2020.

- [30] C. Michel, "Cuneiform Fakes: A Long History from Antiquity to the Present Day," De Gruyter, 2020.
- [31] A. Hamed, S. Hamdany, R. R. O. Al-nima, and L. H. Albak, "Translating Cuneiform Symbols Using Artificial Neural Network," TELKOMNIKA (Telecommunication Computing Electronics and Control), vol. 19, no. 2, pp. 438–443, 2021.
- [32] A. Chapman, E. Simperl, L. Koesten, G. Konstantinidis, L. D. Ibáñez, E. Kacprzak, and P. Groth, "Dataset Search: A Survey," The VLDB Journal, vol. 29, no. 1, pp. 251-272, 2020, Springer.
- [33] T. Jauhiainen, H. Jauhiainen, T. Alstola, and K. Lindén, "Language and dialect identification of cuneiform texts," arXiv preprint arXiv:1903.01891, 2019.
- [34] M. L. Paiva, "SAAO (State Archives of Assyria online)," Journal of Ancient History, vol. 2, pp. 305-310, 2011.
- [35] W. S. Exploration, "Web-Based Scientific Exploration and Analysis Collaborative Research," Informatics. Vol. 4. No. 4. MDPI, 2017.
- [36] P. Taracha, "Hittitology Up to Date : Issues and New Approaches Magical Rituals Known from the Archives of the Hittite Capital Hattusa, Minor in the Second Millennium BC , Hittitology as this Field is Called, has Come of Age 150 km as the Crow Flies ast of Ankara in 1906 , Discovered Over the Course of Four," 2012.
- [37] J. Green and M. W. Stolper, "Achemenet Project," The Oriental Institute , vol. 2014, no. April 2014, pp. 11–13, 2015.
- [38] Fetaya, E., Lifshitz, Y., Aaron, E., & Gordin, S. "Restoration of Fragmentary Babylonian Texts Using Recurrent Neural Networks," Proceedings of the National Academy of Sciences 117.37 (2020): 22743-22751.