

Data Mining Ancient Scripts to Investigate their Relationships and Origins

Shruti Daggumati
University of Nebraska-Lincoln
Lincoln, Nebraska, USA
sdagguma@cse.unl.edu

Peter Z. Revesz
University of Nebraska-Lincoln
Lincoln, Nebraska, USA
revesz@cse.unl.edu

ABSTRACT

This paper describes a data mining study of a set of ancient scripts in order to discover their relationships, including their possible common origin from a single root script. The data mining uses convolutional neural networks and support vector machines to find the degree of visual similarity between pairs of symbols in eight different ancient scripts. Among the surprising results of the data mining are the following: (1) the Indus Valley Script is visually closest to Sumerian pictographs, and (2) the Linear B script is visually closest to the Cretan Hieroglyphic script.

CCS CONCEPTS

• **Computing methodologies** → *Machine learning; Machine learning approaches*; • **Information Systems** → *Data mining*; • **Human-centered computing** → *Visualization; Treemaps*.

KEYWORDS

Convolutional Neural Networks, Data Analytics, Data Mining, Indus Valley Script, Sumerian Pictographs, Support Vector Machines, Data Visualization, Machine Learning

ACM Reference Format:

Shruti Daggumati and Peter Z. Revesz. 2019. Data Mining Ancient Scripts to Investigate their Relationships and Origins. In *23rd International Database Engineering & Applications Symposium (IDEAS'19), June 10–12, 2019, Athens, Greece*. ACM, New York, NY, USA, 10 pages. <https://doi.org/10.1145/3331076.3331116>

1 INTRODUCTION

The data mining work in this paper is motivated to help decipher ancient scripts such as the still undeciphered Indus Valley Script [24]. The idea is that if an undeciphered script can be matched with an already deciphered script, then the phonetic values of the symbols in the deciphered script can be reasonably expected to match the phonetic values of the corresponding symbols in the undeciphered script.

We applied various data mining methods to compare and analyze the relationship among the following ancient scripts: Brahmi, Cretan Hieroglyphs, Greek, Indus Valley, Linear B, Phoenician,

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
IDEAS'19, June 10–12, 2019, Athens, Greece

© 2019 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-1-4503-6249-8/19/06...\$15.00
<https://doi.org/10.1145/3331076.3331116>

Proto-Elamite, and Sumerian Pictographs. Our data mining yields a script family tree with a common origin of all these scripts. A particularly interesting finding of our data mining is that the Indus Valley Script seems to derive from the Sumerian Pictographs. Our finding is supported by the following observations of other authors. First, it is known that intensive trade existed, mainly by sea between the ancient civilizations in Mesopotamia and the Indus valley, and the urbanization, irrigation technology, social organization, commercial patterns, and numerous other features of the Indus Valley civilization bears a close resemblance to the Sumerian model [4, 9]. Second, the ancient Sumerian records referred to the Indus Valley Civilization as *Meluhha*, which means “high country” in Dravidian languages according to Parpola [24] and may be related to the present day region of *Baluchistan*.

The rest of this paper is organized as follows. Section 2 describes the dataset of the ancient scripts and texts which we used as a data source. Section 3 describes the machine learning methodologies that we used for the computerized comparison of the visual characteristics of pairs of symbols from the different scripts. Section 4 presents the experiments and results and analyzes the findings. Section 5 discusses related work. Finally, Section 6 gives some conclusions and directions for further research.

2 DATASET

In this section, we provide the historical background for all the scripts used in this work. We also describe how the datasets were created for the computations.

2.1 Brief Review of the Eight Scripts Considered

2.1.1 Brahmi. Brahmi is the second oldest South Asian script, after the Indus Valley Script. The Brahmi script is an abugida, which uses a system of diacritical marks to denote vowel association with the consonant symbols. The direction of writing for the Brahmi script is left to right. Much like the Indus Valley Script, the Brahmi script has a debated origin.

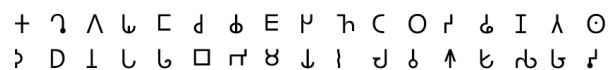


Figure 1: Sample Brahmi script symbols.

2.1.2 Cretan Hieroglyphs. Cretan Hieroglyphs was the first writing of the Minoans and predecessor to Linear A, which in turn gave rise to Linear B and Cypriot. It was used between 2100 to 1700 BC [2, 23]. The second author proposed recently a decipherment of Cretan Hieroglyphs [36], but there are many alternative proposals.



Figure 2: Sample Cretan Hieroglyphs.

2.1.3 *Greek*. There were many variants of the early Greek alphabet, each suited to a local dialect. Eventually, the Ionian alphabet was adopted in all Greek-speaking states. Ancient Greek is a full (consonants and vowels) alphabet. Greek was written from around 800 BC to the 5th century in both a right-to-left and a boustrophedonic style, but later it transitioned to a left-to-right writing system [5].

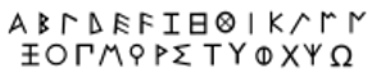


Figure 3: The 26 letters of the ancient Greek alphabet.

2.1.4 *Indus Valley*. The Indus Valley Script is an undeciphered script, which was used between 2400 and 1900 BCE [25]. It is stated to be a logographic and syllabic writing system, written from right to left [25].



Figure 4: Sample Indus Valley script symbols.

2.1.5 *Linear B*. Linear B was used in Mycenaean Greece and is the oldest known Greek writing [15]. Linear B remained a mystery until 1952 when Michael Ventris deciphered Linear B showing that it is an archaic version of Greek [3]. Linear B is a syllabic writing system where in general each syllable begins with a single consonant, which is followed by a single vowel.

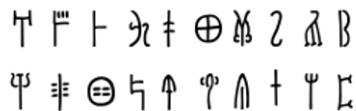


Figure 5: Sample Linear B symbols.

2.1.6 *Phoenician*. The Phoenician alphabet was used from 1200 to 150 BC in the eastern Mediterranean [13]. The Phoenician alphabet is an abjad (only consonants with no vowels) writing system, written from right to left, which consists of 22 letters representing consonants [13]. The Phoenician alphabet may derived from Egyptian Hieroglyphs [16] or Linear B [35].

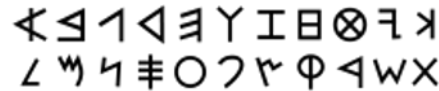


Figure 6: The 22 letters of the Phoenician alphabet.

2.1.7 *Proto-Elamite*. The Proto-Elamite script was briefly used between the end of 4000 to the beginning of 3000 BCE in present-day Iran and southern Iraq [11]. The script uses around 1900 non-numerical signs, although 1700 of those signs only appear a maximum of nine times in the 1600 Proto-Elamite texts [8]. The Proto-Elamite script is said to be logographic or ideographic [11] and is also considered undeciphered.

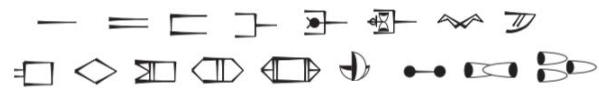


Figure 7: Sample Proto-Elamite script symbols.

2.1.8 *Sumerian Pictographs*. The Sumerian language is distantly related to both the Uralic and the Dravidian language families [28, 41]. However, the Sumerian Pictographs are considered an independent development by most researchers [11]. The Sumerian pictographic script is primarily a syllabic and logographic writing system. It was written from left to right, and it and its cuneiform descendant were used from 3100 BCE to 1st century AD [11].

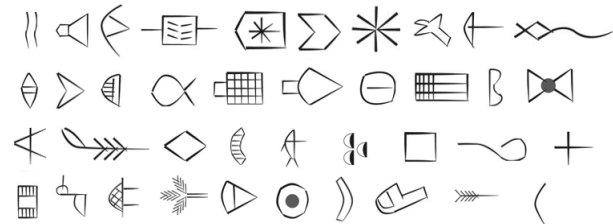


Figure 8: Sample Sumerian Pictographs.

2.2 Data Source

The eight different scripts outlined in the previous section were used as a data source. For the Brahmi script we use 34 of the symbols (Figure 1), for the Cretan Hieroglyphs we use 22 symbols (Figure 2), for Greek we use all 27 symbols (Figure 3). For the Indus Valley Script, we use 23 symbols (Figure 4) which were symbols with the highest frequencies because the Indus Valley Script has at least over 400 symbols and symbols that occur only once or twice are likely to be insignificant [44]. For Linear B we use 20 symbols (Figure 5), for the Phoenician alphabet we use all 22 symbols (Figure 6), for the Proto-Elamite script we use 17 symbols (Figure 7), and for the Sumerian Pictographs we use 34 symbols (Figure 8).

2.3 Data Gathering and Processing

Our dataset is modeled after the MNIST image database [21]. Each symbol in our dataset has 780 training images and 120 validation images, that is a total of nine hundred images associated with each symbol. The images used were hand generated and computer modified via minor skewing and distortion. Each image is 50x50 pixels, grayscale, and centered in the 50x50 region using the center of mass. These features are the necessary preprocessing steps for each dataset.

3 SOFTWARE ARCHITECTURE

3.1 Convolutional Neural Network (CNN)

We created neural networks using Python and TensorFlow with a Keras wrapper. The constructed neural networks have various levels of accuracy, depending on the script learned. The architecture of our convolutional neural network is similar to the LeNet model [20] with a modification on the output classification as shown in Figure 9. The main deviance from the original LeNet model is that we use an SVM classifier for the final dense layer instead of a Softmax layer. Previous works have shown this to be useful in recognition of other languages [10] or even when the sample set is more than ten [21].

Starting with our image size of 50x50 we first apply a convolution using a filter size of 5x5, which reduces our image to 46x46. After this we apply a pooling layer which reduces our image size by half, entailing a 23x23 image. We then add one more convolution layer using a 4x4 filter which reduces our image size to 20x20. Then we apply a pooling layer which reduces our image in half again to 10x10. We then pass the image to a fully connected flattened layer of 1024 neurons, which then passes the data to our SVM (see Section 3.2). Each convolution layer has a Rectified Linear Unit (ReLU) activation function. ReLU is often used as the activation function of choice for most CNN architectures. The ReLU activation function produces zero as an output when $x \leq 0$ or it produces a linear value with slope of one when $x > 0$. Each pooling layer employs max pooling. Each 2x2 filter takes the maximum value of the four quadrants to use for the feature map. To combat overfitting we use a drop rate of 0.4. Each CNN uses the Adam optimizer with learning rate of 0.001. Adam is an adaptive learning rate optimization algorithm that was designed specifically for training deep neural networks [19].

3.2 Support Vector Machine (SVM)

The generated SVM is implemented in Python and uses Python library packages. SVMs were designed for binary classification. However, in our research, we use SVMs for a multiclass problem. Generally, for classification problems in CNNs, the last layer uses Softmax. In this research, we use L2-SVM which is differentiable and optimizes the sum of the squared errors. The L2-SVM also minimizes the squared hinge loss. The optimization function for the L2-SVM is shown below, where w is an N -dimensional weight vector, b is the bias terms, and ξ_i are slack variables, and C is the penalty parameter.

Minimize:

$$\frac{1}{2} \|w\|^2 + \frac{C}{2} \sum_{i=1}^N \xi_i^2 \quad (1)$$

Subject to:

$$y_i(x_i \cdot w + b) \geq 1 - \xi_i \quad i = 1, \dots, N \quad (2)$$

As mentioned previously, training the classifier using the L2-SVM objective function outperforms other methods such as L1-SVM or Softmax regression [48].

3.3 Prediction Classifier

In addition to creating a CNN+SVM classifier per each script, we also look at the similarities between two pairs of scripts. The trained CNN+SVM model for every script is passed into the other seven script models. The basic idea of the predictive classifier is illustrated in Figure 10.

Each similarity matrix produced by the CNN+SVM for the eight scripts has different NxM dimensions based on the number of symbols in each script. We create the following two measures to see the strength between two scripts:

- (1) The **Average of All** takes the average of the strongest probability matches for each symbol pair. The rationale is that taking the average of the strongest matches between two scripts takes into account all the symbols in each script. If a symbol provided as input has a low correlation with all of the trained symbols, the overall average would reflect this.
- (2) The **Selective Average** only considers pairs of symbols which have higher than seventy-five percent similarity match and then take the average. The rationale is that the selective average provides two measures in regards to the similarity of two scripts. It provides not only a higher overall average in comparison to taking the average overall but also the number of symbols which are the closest together. The selective average also takes into account that a script may not completely stem from only one script. Therefore not all symbols may have a high correlation.

3.4 Classification Trees

Each CNN+SVM for the prediction classifier has seven mappings for the eight scripts. The strength between the scripts is provided using the two averages presented in Section 3.3. In addition, we take into account the number of symbols between the scripts which have a correlation value $\geq 75\%$.

To create a classification tree we employ two different algorithms for the two measures as listed below.

- (1) **Similarity:** *The scripts which have a higher correlation are paired.* We use WPGMA (Weighted Pair Group Method with Arithmetic Mean) to create our dendrogram for the scripts. The WPGMA algorithm creates a dendrogram that displays the structure in the similarity matrix. The nearest two clusters are combined at each step i.e. clusters x and y are combined to create $x \cup y$. Then the distance to another cluster z is the mean of the distances between z and $x \cup y$ as shown in Equation (3). Since we use a similarity matrix as input to the WPGMA method, we use the complement of the matrix. That is, now the smaller values indicate higher similarity.

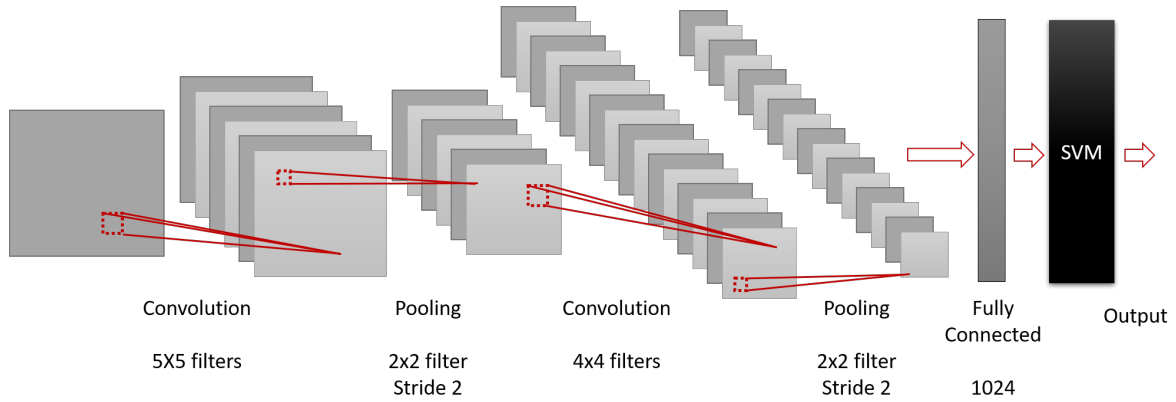


Figure 9: The architecture of our CNN+SVM classifier.



Figure 10: The predictive CNN+SVM classifier comparing the other seven scripts to the Phoenician alphabet. The unknown script is replaced with any of the seven other scripts. The size of the matrix is dependent on the number of symbols in the unknown script provided.

$$d_{(x \cup y), z} = \frac{d_{x,z} + d_{y,z}}{2} \quad (3)$$

- (2) **Hierarchical:** *The scripts are ancestor/descendant of another script.* The hierarchical tree generation is implemented again using WPGMA but also considering the time period when each script was used. By doing this we can create a descendant tree, which highlights the possible descendant of each script. The details are shown below in Algorithm 1.

Algorithm 1 Time-Based Descendant Tree

- 1: Create parent node P
 - 2: Create a node for each script
 - 3: **for all** Closest Script Pairs S_x and S_y **do**
 - 4: **if** $S_x.Time > S_y.Time$ **then**
 - 5: Parent of S_x is P
 - 6: Parent of S_y is S_x
 - 7: **else**
 - 8: Parent of S_y is P
 - 9: Parent of S_x is S_y
 - 10: **for all** Singleton Scripts S_z **do**
 - 11: Parent of S_z is P
 - return** Tree
-

4 EXPERIMENTAL RESULTS AND ANALYSIS

In this paper, we have three main fundamental building blocks: the dataset creation, the CNN+SVM classifier, and the hierarchical tree creator. The latter two portions were first independently verified and then combined to create a final product.

4.1 Validation of the Script Classifier

Each script has its own CNN+SVM classifier. The accuracy of the different scripts is shown in Table 1 with an increase of epochs (step size = 25). We see that for all the scripts at 25 epochs we have already reached the 90% accuracy, similarly to MNIST CNNs.

4.1.1 Script Prediction. For each script, its CNN+SVM classifier has an almost perfect accuracy at 100 epochs. Due to that, we see whether the CNN+SVM can be used to find ancestors and/or descendants of other scripts. We partition this experiment into two categories: **Known Origin** and **Unknown Origin**. The known origin scripts validate our framework and ensure that our tool is capable of reproducing established results. Some specific categorizations:

- (1) **Known Origin:** Phoenician is the ancestor of ancient Greek, as mentioned already by Herodotus, and Brahmi, via Aramaic. Cretan Hieroglyphic script an ancestor of Linear B.
- (2) **Unknown Origin:** The Sumerian Pictographs, the Indus Valley, and the Proto-Elamite scripts have unknown ancestors and descendants.

Table 1: Validation Accuracy

	Number of Epochs			
	25	50	75	100
Brahmi	95.09	98.15	98.24	99.35
Cretan Hieroglyphs	91.09	92.84	94.47	97.53
Greek	93.49	96.26	97.23	98.63
Indus Valley	93.50	95.70	96.85	98.23
Linear B	91.19	93.15	96.42	99.48
Phoenician	93.18	94.77	95.36	97.52
Proto-Elamite	91.93	94.55	97.05	99.09
Sumerian Pictographs	90.79	93.21	96.94	97.40

4.1.2 Validation of Our Method - Known Script Prediction. By using the prediction techniques we aim to see the similarities between the scripts. We first validate our thoughts by passing Greek into the trained Phoenician CNN-SVM and vice-versa. Similarly, we repeat this experiment with Linear B and the Cretan Hieroglyphs.

As seen in Figures 11 and 12, the heatmaps of the similarity matrices between Phoenician and Greek indicates high correlation on the diagonal. This indicates the Phoenician and Greek have an almost one-to-one mapping. We see that this result is validated by the known mapping between Greek to Phoenician as shown in Table 2. We find similar results with Linear B and Cretan Hieroglyphs, which also indicates that the Cretan Hieroglyphs and Linear B have an almost one-to-one mapping.

4.1.3 Unknown Origin Script Prediction. Since the CNN+SVM predictor worked well on the known origin scripts, yielding the expected ancestor-descendant relationships, we can safely use it for the unknown origin scripts too. As visualized in Figure 13, the Sumerian Pictographs and the Indus Valley script have a fairly strong correlation and an almost one-to-one mapping similar to the relation between Phoenician and Greek and between Cretan Hieroglyphs and Linear B. Table 3 notes the number of symbols which have a $\geq 75\%$ correlation between scripts.























4.2 Tree Visualization Analysis

The similarity matrices shown in the previous sections produce the classification and hierarchy trees as shown in Figures 14 and 15, respectively.

4.2.1 Classification Tree. Beside confirming the known origins noted earlier, the classification tree generated some interesting new results. In particular, Brahmi is closest to Phoenician and Greek. The visualization also shows that Brahmi, the Cretan Hieroglyphs, Greek, and Linear B and Phoenician form one branch of the classification tree, while Sumerian Pictographs are closest related to the Indus Valley script.

4.2.2 Hierarchy Tree. The hierarchical tree not only shows the similarity between two pairs of scripts but also visualizes that Greek is a descendant of Phoenician and Linear B is a descendant of Cretan Hieroglyphs. In addition, the Indus Valley script has been classified as a possible descendent of the Sumerian Pictographs. Brahmi and Proto-Elamite have an unknown ancestor. However, they have some

Table 2: Mapping between Greek and Phoenician.

Phoenician		Greek	
	aleph	Α	alpha
	beth	Β	beta
	giml	Γ	gamma
	daleth	Δ	delta
	he	Ε	epsilon
	waw	Ϝ or Υ	digamma or upsilon
	zayin	Ζ	zeta
	heth	Η	eta
	teth	Θ	theta
	yodh	Ι	iota
	kaph	Κ	kappa
	lamedh	Λ	lambda
	mem	Μ	mu
	nun	Ν	nu
	samekh	Ξ	xi
	ayin	Ο	omicron
	pe	Π	pi
	sade	Σ	san
	qoph	Ϟ	koppa
	res	Ρ	rho
	sin	Σ	sigma
	taw	Τ	tau
-		Φ	phi
-		Χ	chi
-		Υ	psi
-		Ω	omega

similarities to the other scripts to assume an unknown hypothetical common origin of these eight scripts.

5 RELATED WORK

5.1 Background - Indus Valley Script

Sir Alexander Cunningham, one of the first to encounter the Indus Valley script, assumed that the seals were foreign import. He later stated that Brahmi might be a descendant of the Indus Valley script.

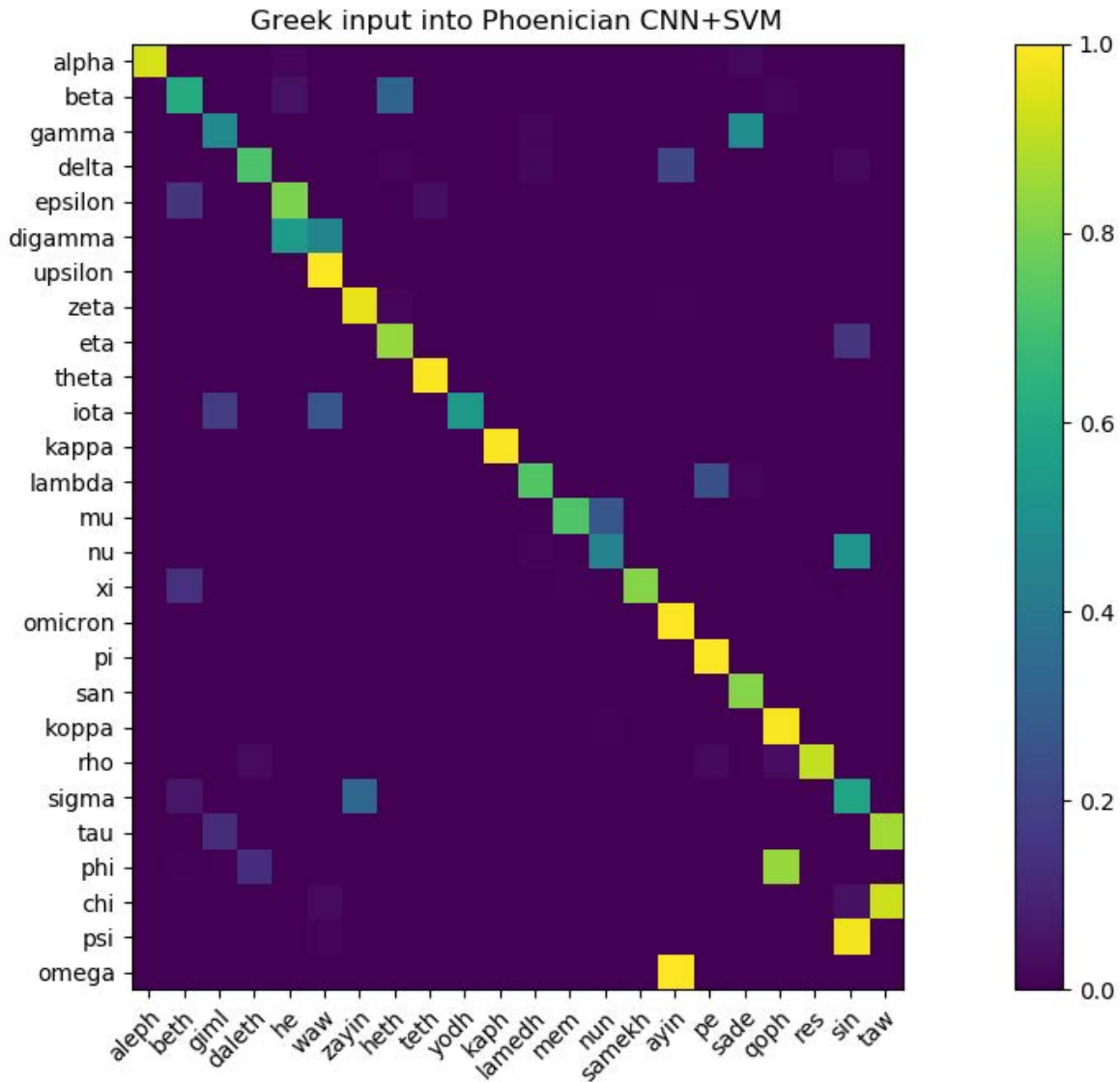


Figure 11: The Greek letters are provided as input to the trained Phoenician CNN+SVM.

Table 3: The number of symbols with correlation $\geq 75\%$ between each pair of the eight scripts.

	Brahmi	Cretan Hier.	Greek	Indus Valley	Linear B	Phoenician	Proto-Elamite	Sumerian Pict.
Brahmi	34	-	-	-	-	-	-	-
Cretan Hier.	2	22	-	-	-	-	-	-
Greek	9	4	26	-	-	-	-	-
Indus Valley	8	5	9	23	-	-	-	-
Linear B	3	20	7	4	20	-	-	-
Phoenician	9	6	22	9	9	22	-	-
Proto-Elamite	2	2	2	4	0	3	17	-
Sumerian Pict.	6	6	7	20	5	7	3	39

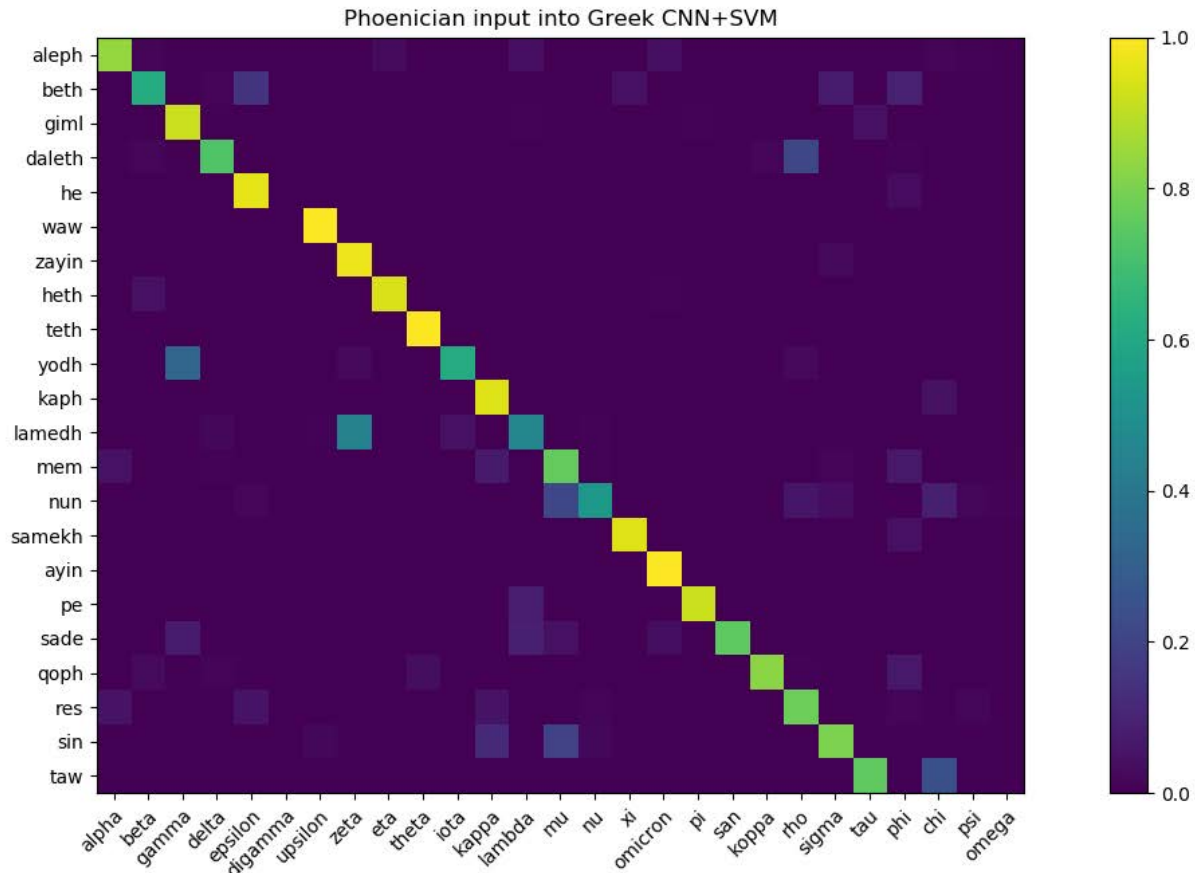


Figure 12: The Phoenician letters are provided as input to the trained Greek CNN+SVM.

Many other scholars have connected the Indus Valley script to Brahmi [29–31]. Many scholars also suppose that the Indus Valley Script expresses some Dravidian language [24, 26, 27, 43, 45, 46, 49], where the work from [49] was one of the first publications using computer aid to analyze the Indus Valley script.

Some scholars, such as McAlpin [22], support the Elamo-Dravidian hypothesis, which links the Dravidian to the Elamite languages. McAlpin also believes that the Indus Valley script could be part of the Elamo-Dravidian language family. That hypothesis is supported by evidence of extensive trade between Elam and the Indus Valley civilization.

There are a few scholars who believe that the Indus Valley script is not a language [12]. These scholars say that the Indus Valley script is comparable to nonlinguistic signs which symbolize family or clan names/symbols and religious figures/concepts. Regardless of it being a language or not, its similarity to the other scripts still suggests that the symbols were derived from Sumerian Pictographs.

Nevertheless, the brevity of Indus texts may indeed suggest that it represented only limited aspects of an Indus language. That is true of the earliest, proto-cuneiform, writing on clay tablets from

Mesopotamia, around 3300 BC, where the symbols record only calculations with various products (such as barley) and the names of officials.

5.2 Machine Learning

Scholars have used various machine learning techniques to analyze and classify images and read text [17, 18].

Support vector machines and neural networks have been used to recognize a multitude of scripts. Artificial neural networks and SVMs were compared on the Devanagari script, a descendant of the Brahmi script [1]. Arabic handwritten recognition was recently studied using the CNN+SVM combination [10]. In addition, handwritten Chinese characters were analyzed using CNNs [14, 47]. Earlier work of the authors shows the similarity between the Indus Valley script and other scripts using CNNs [6, 7]. However, the use of neural networks to generate script families is a new domain.

5.3 Classification Trees

Revesz [34, 35] used hypothetical evolutionary tree reconstruction algorithms to analyze the development of the Cretan Script Family.

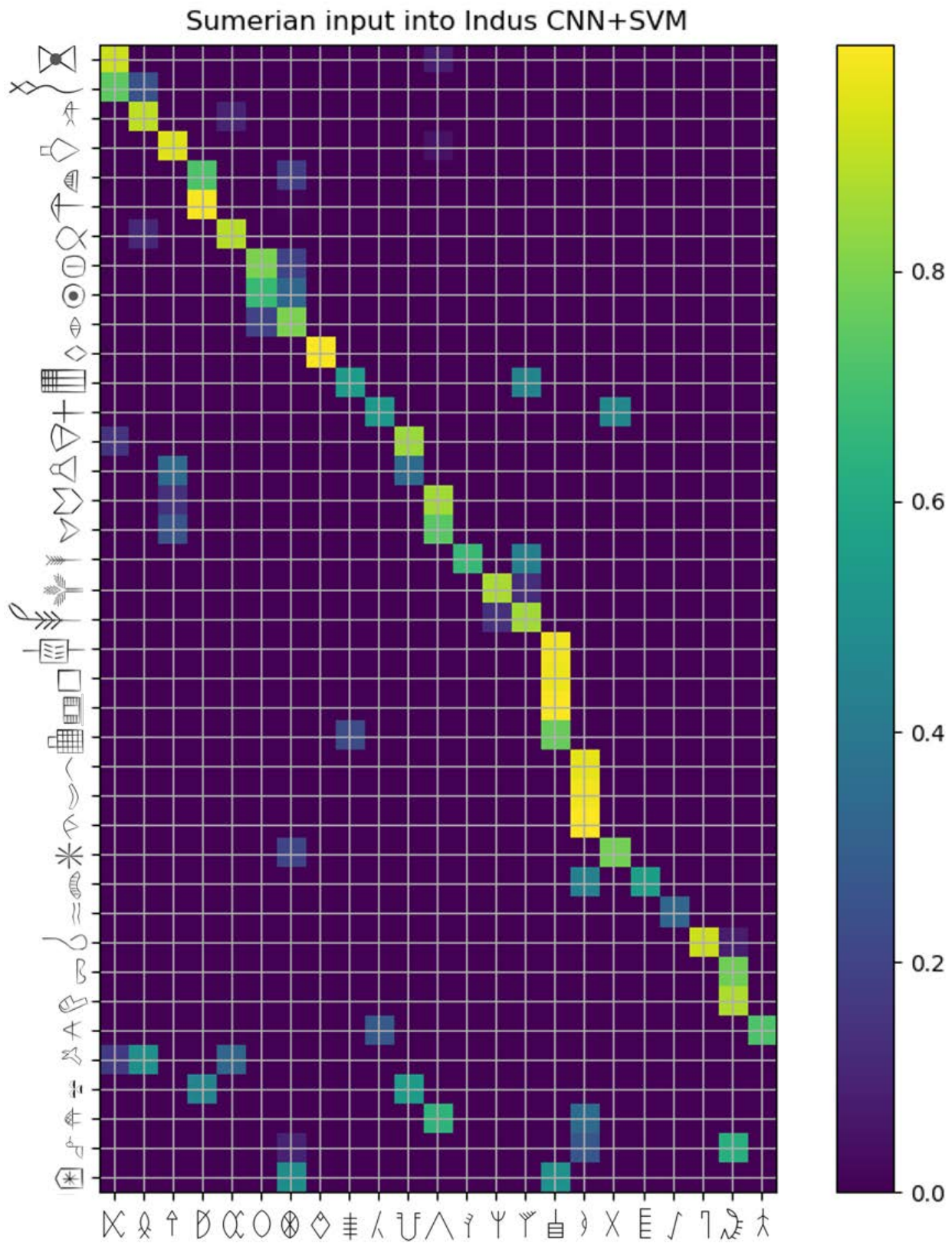


Figure 13: The Sumerian Pictograms are provided as input to the trained Indus Valley CNN+SVM.

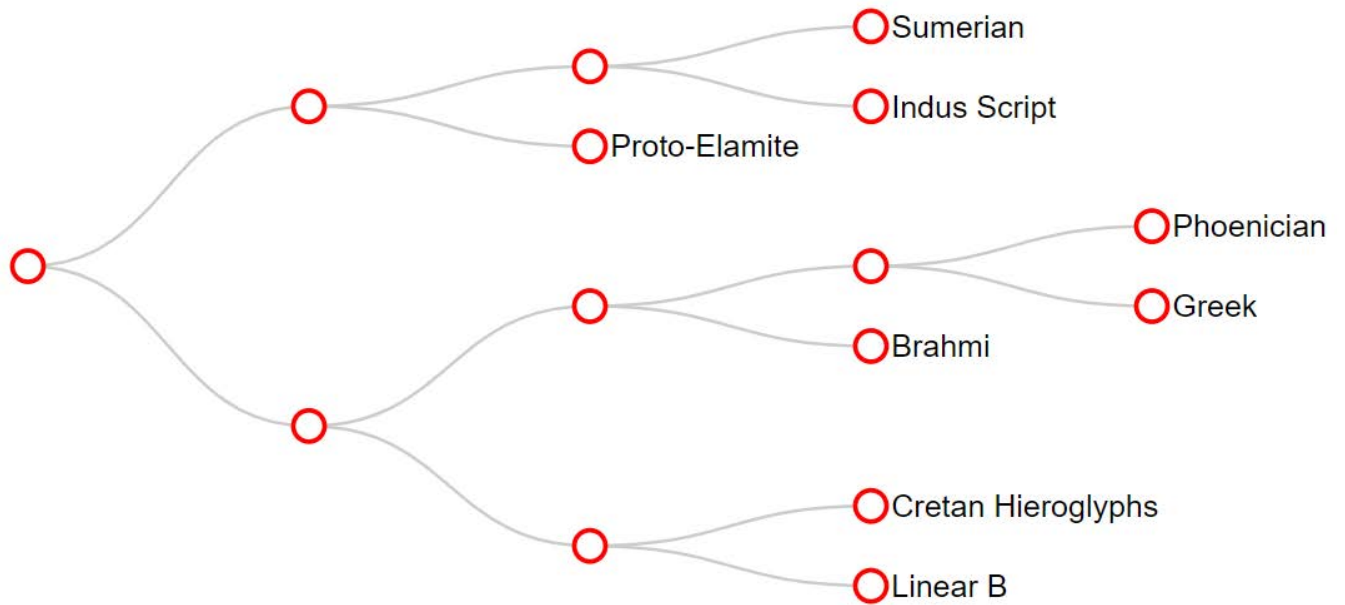


Figure 14: The classification tree created from the similarity matrix using WPGMA.

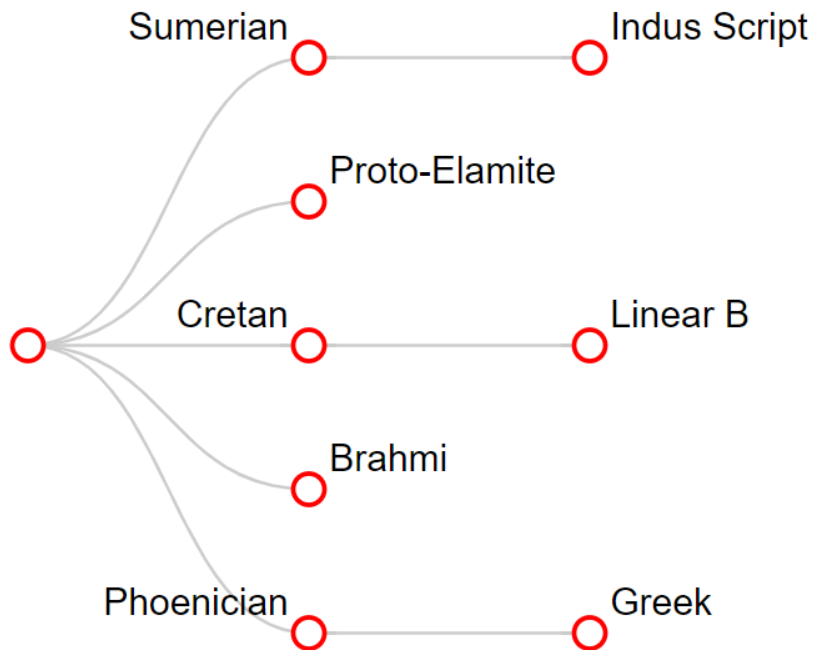


Figure 15: The hierarchical tree created by taking time into account.

The matching of Minoan Cretan Hieroglyphic and Linear A symbols with the Carian and the Old Hungarian alphabets yielded new phonetic values for the Cretan Hieroglyphic and Linear A symbols. The new phonetic values allowed the decipherment of the Linear A script [39], and the Cretan Hieroglyphic script [36], including the Arkalochori Axe [40] and the Phaistos Disk [37] inscriptions. The AIDA system [42] is an online Minoan inscriptions database that also contains some of these translations.

The origin of languages and scripts have long been studied by linguists. The use of genetic information tying civilizations and their languages have only recently been studied [32, 33, 38]. Using human archaeogenetics may provide new insight into the diffusion of human populations in association with various language families.

6 CONCLUSIONS AND FUTURE WORK

The invention and spread of writing was a giant step for humanity that is still largely shrouded in mystery. However, our data mining of ancient script databases revealed several interesting hitherto unknown relationships among the eight scripts studied. This work is only the beginning of a systematic neural networks-based exploration of an ancient script family that likely encompasses not only the eight scripts that we studied but also many others. Hence as a future work, we plan to add to our database other ancient scripts from the region of the Near East and the Mediterranean Sea. By adding more scripts to our CNN+SVM predictor system, we can obtain a more complete tree of visual similarities and reduce the remaining uncertainties in the development of one of the oldest script families in the world.

REFERENCES

- [1] S. Arora, D. Bhattacharjee, M. Nasipuri, L. Malik, M. Kundu, and D. K. Basu. Performance comparison of SVM and ANN for handwritten Devnagari character recognition. *arXiv preprint arXiv:1006.5902*, 2010.
- [2] J. G. P. Best and F. C. Woudhuizen, editors. *Ancient Scripts from Crete and Cyprus*, volume 9. Bill Archive, 1988.
- [3] J. Chadwick. *The Decipherment of Linear B*. Cambridge University Press, 1958.
- [4] D. Collon. Mesopotamia and the Indus: The evidence of the seals. In *The Indian Ocean in Antiquity*, pages 209–225. The British Museum and Kegan Paul International London/New York, 1996.
- [5] B. F. Cook. *Greek Inscriptions*, volume 5. University of California Press, 1987.
- [6] S. Daggumati. Similarity queries on script image databases. In A. Benczur, B. Thalheim, T. Horváth, S. Chiusano, T. Cerquitelli, C. I. Sidló, and P. Z. Revesz, editors, *New Trends in Databases and Information Systems - ADBIS 2018 Short Papers and Workshops*, pages 391–401. Springer, 2018.
- [7] S. Daggumati and P. Z. Revesz. Data mining ancient script image data using convolutional neural networks. In *Proceedings of the 22nd International Database Engineering and Applications Symposium*, pages 267–272. ACM, 2018.
- [8] J. L. Dahl. Complex graphemes in Proto-Elamite. *Cuneiform Digital Library Journal*, 2005(6), 2005.
- [9] C. Elisabeth and D. Caspers. Sumer, coastal Arabia and the Indus Valley in protoliterate and early dynastic eras: Supporting evidence for a cultural linkage. *Journal of the Economic and Social History of the Orient/Journal de l'histoire economique et sociale de l'Orient*, pages 121–135, 1979.
- [10] M. Elleuch, N. Tagougui, and M. Kherallah. A novel architecture of CNN based on SVM classifier for recognizing Arabic handwritten script. *International Journal of Intelligent Systems Technologies and Applications*, 15(4):323–340, 2016.
- [11] R. K. Englund. The Proto-Elamite script. In P. T. Daniels and W. Bright, editors, *The World's Writing Systems*, pages 160–164. Oxford University Press, 1996.
- [12] S. Farmer, R. Sproat, and M. Witzel. The collapse of the Indus-script thesis: The myth of a literate Harappan civilization. *Electronic Journal of Vedic Studies*, 11(2):19–57, 2016.
- [13] S. R. Fischer. *History of Writing*. Reaktion Books, 2004.
- [14] M. He, S. Zhang, H. Mao, and L. Jin. Recognition confidence analysis of handwritten Chinese character with CNN. In *Proceedings of the 13th International Conference on Document Analysis and Recognition*, pages 61–65. IEEE, 2015.
- [15] J. T. Hooker and J. H. Betts. *Linear B: An Introduction*. Bristol Classical Press, Bristol, UK, 1980.
- [16] M. C. Howard. *Transnationalism in Ancient and Medieval Societies: The Role of Cross-Border Trade and Travel*. McFarland, 2014.
- [17] M. Jaderberg, K. Simonyan, A. Vedaldi, and A. Zisserman. Reading text in the wild with convolutional neural networks. *International Journal of Computer Vision*, 116(1):1–20, 2016.
- [18] K. Jung, K. I. Kim, and A. K. Jain. Text information extraction in images and video: A survey. *Pattern Recognition*, 37(5):977–997, 2004.
- [19] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference on Learning Representations*, 2015.
- [20] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, et al. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.
- [21] Y. LeCun, C. Cortes, and C. Burges. MNIST handwritten digit database. <http://yann.lecun.com/exdb/mnist/>, 1998.
- [22] D. W. McAlpin. Proto-Elamo-Dravidian: The evidence and its implications. *Transactions of the American Philosophical Society*, 71(3):1–155, 1981.
- [23] J.-P. Olivier. Cretan writing in the second millennium BC. *World Archaeology*, 17(3):377–389, 1986.
- [24] A. Parpola. The Indus script: A challenging puzzle. *World Archaeology*, 17(3):399–419, 1986.
- [25] A. Parpola. The Indus Script. In P. T. Daniels and W. Bright, editors, *The World's Writing Systems*, pages 165–171. Oxford University Press, 1996.
- [26] A. Parpola. Study of the Indus script. In *Proceedings of the International Conference of Eastern Studies*, volume 50, pages 28–66, 2005.
- [27] A. Parpola. *Deciphering the Indus script*. Cambridge University Press, 2009.
- [28] S. Parpola. *Etymological Dictionary of the Sumerian Language*, volume 1 and 2. Foundations for Finnish Assyriological Research, Helsinki, Finland, 2016.
- [29] R. P. Rao, N. Yadav, M. N. Vahia, H. Joglekar, R. Adhikari, and I. Mahadevan. Entropic evidence for linguistic structure in the Indus script. *Science*, 324(5931):1165–1165, 2009.
- [30] R. P. Rao, N. Yadav, M. N. Vahia, H. Joglekar, R. Adhikari, and I. Mahadevan. A Markov model of the Indus Script. *Proceedings of the National Academy of Sciences*, 106(33):13685–13690, 2009.
- [31] S. R. Rao. *The Decipherment of the Indus Script*. Asia Publishing House, 1982.
- [32] C. Renfrew. Archaeology, genetics and linguistic diversity. *Man*, pages 445–478, 1992.
- [33] P. Z. Revesz. *Introduction to Databases: From Biological to Spatio-Temporal*. Springer, 2010.
- [34] P. Z. Revesz. An algorithm for constructing hypothetical evolutionary trees using common mutations similarity matrices. In *Proc. 4th ACM International Conference on Bioinformatics and Computational Biology (ACM BCB)*, pages 731–734, 2013.
- [35] P. Z. Revesz. Bioinformatics evolutionary tree algorithms reveal the history of the Cretan Script Family. *International Journal of Applied Mathematics and Informatics*, 10:67–76, 2016.
- [36] P. Z. Revesz. A computer-aided translation of the Cretan Hieroglyph script. *International Journal of Signal Processing*, 1:127–133, 2016.
- [37] P. Z. Revesz. A computer-aided translation of the Phaistos Disk. *International Journal of Computers*, 10:94–100, 2016.
- [38] P. Z. Revesz. A mitochondrial DNA-based model of the spread of human populations. *International Journal of Biology and Biomedical Engineering*, 10:124–133, 2016.
- [39] P. Z. Revesz. Establishing the West-Ugric language family with Minoan, Hattic and Hungarian by a decipherment of Linear A. *WSEAS Transactions on Information Science and Applications*, 14:306–335, 2017.
- [40] P. Z. Revesz. A translation of the Arkalochori Axe and the Malia Altar Stone. *WSEAS Transactions on Information Science and Applications*, 14(1):124–133, 2017.
- [41] P. Z. Revesz. Sumerian contains Dravidian and Uralic substrates associated with the Emegir and Emesal dialects. *WSEAS Transactions on Information Science and Applications*, 16(1):8–30, 2019.
- [42] P. Z. Revesz, M. P. Rashid, and Y. Tuyishime. The design and implementation of AIDA: Ancient Inscription Database and Analytics system. In *Proceedings of the 23rd International Database Engineering and Applications Symposium*, 2019.
- [43] B. Wells. *An introduction to Indus writing*. University of Calgary, 1998.
- [44] B. Wells and A. Fuls. Online Indus Writing Database. <http://caddy.igg.tu-berlin.de/indus/welcome.htm>, 2017.
- [45] B. K. Wells. *Epigraphic Approaches to Indus Writing*. Oxbow Books, 2011.
- [46] B. K. Wells and A. Fuls. *The Archaeology and Epigraphy of Indus Writing*. Archaeopress, 2015.
- [47] W. Yang, L. Jin, and M. Liu. Chinese character-level writer identification using path signature feature, DropStroke and deep CNN. In *Proceedings of the 13th International Conference on Document Analysis and Recognition*, pages 546–550. IEEE, 2015.
- [48] M. L. Yann and Y. Tang. Learning deep convolutional neural networks for X-ray protein crystallization image analysis. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, pages 1373–1379. AAAI Press, 2016.
- [49] A. R. Zide and K. V. Zvelebil. *The Soviet Decipherment of the Indus Valley Script: Translation and Critique*, volume 156. Walter de Gruyter, 1976.