Journal of Computer Applications in Archaeology

Machine Learning Applications in Use-Wear Analysis: A Critical Review

REVIEW

ANASTASIA ELEFTHERIADOU
SHANNON P. MCPHERRON
JOÃO MARREIROS

*Author affiliations can be found in the back matter of this article

ubiquity press

ABSTRACT

Use-wear analysis examines the macroscopic and microscopic patterns of traces left on tool surfaces as a result of use. Recently, machine learning (ML) has been employed as a promising method for automating and standardizing the identification of these traces. While the number of use-wear analysts using ML continues to grow, discussions regarding the effectiveness and appropriate implementation of these methods are ongoing. The main aim of this literature review is to provide recommendations for the more effective application of ML in use-wear analysis and archaeological research, by identifying trends, research gaps, and evaluating the quality of the models developed.

There are three key challenges identified. Firstly, the limited adoption of open science practices restricts the creation of large datasets and hinders reproducibility and transparency. Secondly, research efforts are concentrated within limited institutions, focusing on certain research questions, algorithms, raw materials, and use-wear traces. Thirdly, the inadequate quality, quantity, and diversity of data affect the performance of the models being developed.

To address these challenges, this paper advocates for the promotion of open science and the systematic gathering of experimental and analytical data. Involving a broader range of institutions can improve research quality and promote greater diversity of perspectives. Collaboration with computer scientists and computational archaeologists is essential to integrate the expertise necessary for designing and implementing effective ML methods. By addressing these factors, this paper facilitates the effective use of machine learning, enabling use-wear analysts and archaeologists to develop robust models that automate, accelerate, and improve their research.

CORRESPONDING AUTHOR:
Anastasia Eleftheriadou
ICArEHB, University of Algarve, PT
anasiagre@hotmail.com

KEYWORDS:
use-wear analysis; machine learning; computational archaeology; open science; FAIR data

TO CITE THIS ARTICLE:
Eleftheriadou, A, McPherron, SP and Marreiros, J. 2025. Machine Learning Applications in Use-Wear Analysis: A Critical Review. *Journal of Computer Applications in Archaeology*, 8(1): 188–205. DOI: https://doi.org/10.5334/jcaa.190

1. INTRODUCTION

Lithic artifacts play a pivotal role in the study of human evolution, as changes in their technology and typology are closely tied to the emergence of fundamental social behaviors (Ambrose 2001; Caneva 2001; Foley & Lahr 2003; Kuhn 2020; Lycett 2015; Režek et al. 2018). The variability observed in lithic, bone and wooden artifacts may be indicative of adaptive strategies shaped by environmental, demographic, and social factors influencing tool production and use (e.g., Foley & Lahr 2003; Hovers & Belfer-Cohen 2013; Lemorini et al. 2014; Shea 2017). Use-wear analysis, combined with experimental archaeology, provides further insight into tool use by establishing reference libraries against which archaeological artifacts are compared (Evans 2014; Fullagar 2014; Hayden 1979; Keeley 1980; Marreiros et al. 2015; Marreiros, Pereira & Iovita 2020). Based on experimental replications, diagnostic macro- and micro-traces of use on bone and lithic surfaces are known to correlate with specific worked materials and motions (Marreiros et al. 2020). Bone surface modifications (BSM) are similar to use-wear traces but specifically apply to bone specimens, documenting alterations caused by humans, animals, or other natural factors, providing insights into taphonomy and subsistence patterns (Fisher 1995).

During the last decades, several studies started implementing methods that allow a quantitative measurement of use-wear features (Borel et al. 2021; Evans & MacDonald 2011). These methods include the use of new imaging and metrology techniques, such as confocal microscopy, focus variation microscopy, interferometry, and atomic force microscopy for the acquisition and processing of 3D surface texture data (Evans & Donahue 2008; Faulks et al. 2011). More recently, artificial intelligence (AI) has been used as a novel approach that combines both qualitative and quantitative data, reduces time and user-based bias, as well as increases efficiency in the characterization and analysis of use-wear traces (Zhang et al. 2024).

AI is a broad field that encompasses various subfields, and its definition can vary depending on context (Samoili et al. 2020). In this article, we adopt the definition from Eurostat where 'Artificial intelligence refers to systems that use technologies such as: text mining, computer vision, speech recognition, natural language generation, machine learning, deep learning to gather and/or use data to predict, recommend or decide, with varying levels of autonomy, the best action to achieve specific goals' (Samoili et al. 2020: 9).

In the field of use-wear analysis, AI was first implemented in a doctoral thesis using expert systems (van den Dries 1998). Expert systems mimic the decision-making abilities of human experts and are used for tasks such as student education, data analysis, and hypothesis validation (van den Dries 1998). Two interactive and user-friendly applications were developed for this purpose. The Wear Analyzing and Visualizing Expert System (WAVES) was the first application used for analysis, guiding the process of use-wear identification and interpretation, and for evaluating interpretations made by more experienced students or analysts (ibid.). The second application was the Wear Analysis and Recognition Neural Network Prototype (WARP), which used wear attributes provided by the user to relate wear traces (specifically polish) to contact materials (ibid.).

Following the work by van den Dries, subsequent research on use-wear analysis began to focus on Machine Learning (ML), a subset of AI with the capacity to learn and improve its performance through the use of computational algorithms (Bini 2018). There are two types of ML: supervised and unsupervised. These differ in the type of data they handle and the application they are used for (Raschka & Mirjalili 2019). Supervised learning uses labelled data (raw data annotated with labels) as input to build a model that can then make predictions regarding new, previously unseen data (Yravedra et al. 2021; Zhang et al. 2024). Unsupervised learning uses unlabeled data (raw data without annotations) as input to explore the structure of the data and extract meaningful information (Courtenay et al. 2020a; Yezzi-Woodley et al. 2022). Among the most commonly used algorithms in use-wear analysis are neural networks and deep learning, both of which are subsets of ML. Neural networks (NN) process data like the human brain using a single network of interconnected nodes, called neurons, that accept, store, and pass information to process input data to produce an output (Bini 2018). Deep learning (DL) uses NNs with multiple layers to learn hierarchical representations of data (Kubat 2017). Depending on the architecture and type of data used, there are different types of DL algorithms, such as convolutional neural networks optimized for image data and recurrent neural networks optimized for sequential data (e.g., time series and text; Raschka & Mirjalili 2019). The training process of both NN and DL models relies on backpropagation, an optimization method that adjusts the model's parameters during training to minimize error and improve prediction accuracy (Bishop 2006). Examples of studies using DL in use-wear analysis include Luncz et al. (2022), who classified damaged versus undamaged surface structures on wooden tools, and Zhang et al. (2024), who classified polish based on contact material. In the context of BSMs, studies employing DL include Abellán et al. (2021), who classified carnivore tooth scores, and Courtenay et al. (2020b), who classified cut marks and trampling marks.

A common challenge in developing NN and DL models is the substantial amount of data required for training to increase the generalization of models (i.e., how well the model performs on unseen data; Merchant & Castleman 2022). The definition of a 'minimum dataset size' varies depending on several factors, including the

model's complexity and type, number of predictors, and scientific domain (Xu et al. 2023; Zantvoort et al. 2024). For this review, we defined a small dataset as one with fewer than 200 samples. This threshold aligns with the definition of 'small' datasets in other archaeological studies that have applied DL, where datasets such as 306 annotated tumuli samples for remote sensing (Berganzo-Besga et al. 2021), 150 polish samples for use-wear analysis (Sferrazza 2025), and 216 tooth pit and score samples for bone surface modification (Courtenay et al. 2019) were considered small. Transfer learning, introduced by Yosinski et al. (2014), addresses this issue by reusing weights from related tasks, thus reducing data, computing power, and time needs (Goodfellow, Bengio & Courville 2016; Merchant & Castleman 2022). Its robustness as a feature extractor and classifier has also been demonstrated in use-wear studies (e.g., Pizarro-Monzo et al. 2022; Zhang et al. 2024).

To date, two studies have evaluated the application of ML in the analysis of micro- and macro-traces of lithic and bone materials. Calder et al. (2022) provide a review of ML applications in archaeology, including use-wear analysis, and highlight key challenges such as ensuring rigorous data partitioning (e.g., train-test splits), cross-validation methods, and reproducibility through data and code sharing. Similarly, McPherron et al. (2022) re-examined a case study on bone surface modifications, emphasizing the importance of sample size, open-science practices, and adherence to methodological standards in ML workflows.

Given that the application of ML in archaeology is a relatively new and rapidly evolving subfield, its methods should be adapted to meet the needs of archaeological datasets and research questions. To address these gaps and challenges, this study synthesizes available use-wear and BSM studies that employ ML, providing targeted observations and recommendations to improve and expand its application. To achieve this aim, three main objectives were defined. The first objective investigates the introduction and evolution of ML in use-wear analysis of lithic and wooden artifacts, as well as BSM, focusing on trends in materials, topics, and algorithms. Despite differences in material types (lithics, bone, and wood) and the distinct research questions associated with use-wear analysis and BSMs, three key factors support a combined review of their ML applications. First, all three material types are analyzed using microscopy to examine macro- and micro-traces, employing comparable image-based and quantitative techniques for data acquisition (Fisher 1995; Marreiros, Pereira & Iovita 2020). Second, they rely on similar types of input data, primarily 2D images or tabular datasets, which enables the use of shared preprocessing techniques, feature extraction methods, and analytical approaches in ML. Third, the limited number of case studies for each material type restricts the ability to assess variability in ML approaches, thereby hindering the identification of material-specific trends. Therefore, while material differences may influence the analysis to some extent, they do not preclude the identification of overarching trends. For a more comprehensive discussion of the archaeological aspects of lithic, bone, and wooden material analysis, readers may refer to Caruso Fermé & Aschero (2020), James & Thompson (2015) and Marreiros, Gibaja Bao & Bicho (2015). The second objective was to examine the variability in dataset size and class balance across the reviewed articles, considering their critical role in influencing model accuracy and generalization (Merchant & Castleman 2022). The third objective was to assess the quality of the models being developed to identify areas requiring improvement and the strategies required to address them. The critical assessment of published case studies aims to assist archaeologists in better understanding how ML can be effectively applied in their field, and to highlight the essential factors for developing robust models that can be used by other researchers for various materials and research questions.

2. MATERIALS AND METHODS

2.1. ARTICLE SELECTION

We selected articles to cover both the introduction and breadth of ML applications in use-wear analysis. Articles applying statistical methods for classification tasks in use-wear and BSM analysis (i.e., Bonney 2014; González-Urquijo & Ibáñez-Estévez 2003; Ibáñez, González-Urquijo & Gibaja 2014; Ibáñez, Lazuen & González-Urquijo 2019; Ibáñez & Mazzucco 2021; Ma et al. 2023) were also included, as they provided the basis for the application of ML (Friedrich et al. 2022). Following Bzdok, Altman & Krzywinski (2018), statistical methods focus on inference, using probability models to understand relationships in the data, while machine learning methods prioritize prediction, using general algorithms to identify patterns in complex and extensive datasets (Table 1).

Our review includes both open-access and restricted-access articles published in English and available online. Google Scholar was the search engine used to identify relevant papers, applying combinations of the keywords 'use-wear analysis', 'machine learning', 'neural networks', 'artificial intelligence', 'bone', 'lithic', 'tool', 'archaeology', 'bone surface modifications' and 'taphonomy'. To expand the search and capture potentially overlooked articles not indexed by Google Scholar, we also used the AI-based scholarly discovery tool ResearchRabbit (Cole & Boutet 2023). All papers satisfying the above criteria were collected, reaching a total of 48 case studies. Nearly all case studies consisted of papers published in peer-reviewed journals, with the exception of one from a non-peer-reviewed source. Recognizing the potential limitations of non-peer-reviewed studies, the paper was examined and interpreted with caution. All identified papers were included in the review, covering the period from 1998 to the present.

GROUP	CATEGORY	SUB-CATEGORY	DESCRIPTION
Machine Learning	Instance-Based Methods	KNN	K-Nearest Neighbor
	Decision Trees	DT	Decision Tree
		CTREE	Conditional Inference Trees
	Ensemble Methods	RF	Random Forest
		GBMAC	Gradient Boosted Machines
		GBMOD	Generalized Boosted Model
	Kernel Methods	SVM	Support Vector Machines
	Probabilistic Methods	NBC	Naïve Bayes Classifier
		GNB	Gaussian Naïve Bayes
	Neural Networks	NN	Neural Networks
		CNN	Convolutional Neural Networks
		DCNN	Deep Convolutional Neural Networks
	Clustering Methods	MS	Mean-Shift
		SLC	Single Linkage Clustering
		DBSCAN	Density-Based Spatial Clustering Algorithm with Noise
Statistics	Regression-Based	LR	Linear Regression
		GLM	Likelihood-Based Generalized Linear Model
		CVLR	Cross-Validated Logistic Regression
	Discriminant Analysis	LDA	Linear Discriminant Analysis
		QFDA	Quadratic Discriminant Function Analysis
		DA	Discriminant Analysis
		PLSDA	Partial Least Squares Discriminant Analysis
		MDA	Mixture Discriminant Analysis
		FDA	Flexible Discriminant Analysis
		SDA	Shrinkage Discriminant Analysis
	Dimensionality Reduction	PLS	Partial Least Squares

Table 1 List of algorithms used in the papers under review organized by the group and category of analysis, along with their abbreviations.

2.2. DATA COLLECTION

Data were collected manually by parsing the documents and entering the relevant information into a spreadsheet. The spreadsheet consists of 28 columns, each allocated to store numerical and categorical data related to distinct parameters, including (a) general information (i.e., authors, publication date and title, affiliation, country, access, type of publication, publisher), (b) use-wear (i.e., type of analysis, raw material), and (c) ML (i.e., sample size, data distribution, data dimensionality, type of ML and the specific algorithm used, type of analysis, use of script or software and its access, evaluation metrics, model accuracy, limitations, training, and validation loss). The limitations identified by the authors of the reviewed papers were grouped into five categories: 'Archaeological constraints' (parameters that affect the preservation and interpretation of traces such as taphonomy and equifinality), 'Methodological issues' (unbalanced or insufficient sample size), 'Data acquisition process' (limitations in creating or analyzing use-wear traces and BSM for ML models), 'Resources' (challenges related to the financial or temporal investment in developing ML models), and 'Restricted open science' (non-adherence to FAIR principles of findability, accessibility, interoperability, and reusability; Wilkinson et al. 2016).

To ensure consistency across the spreadsheet, semicolons were used to separate values, special characters (e.g., ñ) were replaced with basic Latin equivalents and empty cells were populated with 'NA'. Entries consisting of more than one word were either connected with underscores (e.g., cut_marks) or abbreviated to facilitate the subsequent creation of graphs (e.g., 'CTREE' for 'Conditional Inference Trees'; see Table 1). Following data collection, the spreadsheet was converted into a comma-separated value (.csv) file and imported into Python scripts for data post-processing, analysis, and presentation. Descriptive statistics were used to summarize the characteristics and distribution

of the values in our dataset (Starbuck 2023). The basic numerical operations were performed using the NumPy library (version 1.22.4; Harris et al. 2020). The Pandas (v 1.4.3) and Geopandas (v 1.0.1) libraries were used for data manipulation and processing (Jordahl et al. 2020; The pandas development team 2024). The Quarto project with the Python scripts and data for replicating our method and generating all figures is referenced in the Data Accessibility section.

2.3. EVALUATION OF MACHINE LEARNING MODELS

An important component of this review is to assess the quality of ML models developed for use-wear and BSM in archaeology. Evaluating ML requires consideration of several factors, including the quality, quantity and type of data, preprocessing techniques, model selection, hyperparameter tuning, and evaluation metrics. In this study, several of these aspects were examined, focusing specifically on class imbalance, accuracy, loss, dataset size, and overfitting. These criteria were selected to provide a general overview of model performance. However, these metrics do not determine the overall quality of the reviewed papers, nor do they fully capture the broader contributions of the studies, such as methodological innovations, theoretical insights, or practical applications in archaeology. Rather, they serve as quantifiable indicators through which studies can be compared systematically. Additional evaluation criteria, such as underfitting, confusion matrices, and ROC-AUC curves, were not included because of limitations in data availability.

2.3.1. Model stability and robustness

Evaluation metrics quantify an algorithm's performance based on correct (true-positives and true-negatives) and incorrect predictions (false-positives and false-negatives; Handelman et al. 2019). A common metric is the learning curve, which plots either the accuracy (proportion of correct outputs) or loss (error rate or else proportion of incorrect outputs) of a model during training and validation (Raschka & Mirjalili 2019). The learning curve of a robust model should exhibit two important traits. First, it should be 'relatively smooth and monotonically non-decreasing' (Weiss & Tian 2008: 262) without abrupt changes or discontinuities as these may indicate an inappropriate learning rate or an insufficient training dataset (Goodfellow, Bengio & Courville 2016; Viering & Loog 2023). To detect the potential presence of these factors in the papers under review, we quantified the degree of roughness (r) in the loss curves during model validation. The curves were extracted as .png files from the articles, imported into the open-source software QuPath (version 0.5.1; Bankhead et al. 2017), digitized using the polyline tool, and then exported as GeoJSON files. The roughness index (*r*) captures the magnitude and frequency of fluctuations in a learning curve and is analogous to the International Roughness Index (IRI) used in engineering (Sayers 1995). Second, the gap between the training and validation loss curves should be minimal (Raschka & Mirjalili 2019). A large gap is an indication of overfitting, where the model fits the training data too closely and fails to predict previously unseen data (Singh 2019). To quantify the magnitude of the gap between the learning curves, we manually recorded the final loss values for both the training and validation curves across case studies (columns 'max_loss_train' and 'max_loss_val') and compared them using a box plot, with the upper and lower limits representing the final validation and training losses, respectively.

2.3.2. Quantification of class imbalance

A common issue in ML studies is the presence of class imbalance, referring to the phenomenon where different groups (classes) in a training set have an uneven number of examples, thus introducing the risk of poor model performance, inaccurate results and possible bias towards the majority class (Ghosh et al. 2024). Following Buda et al. (2018:251), two equations which characterize the distribution and intensity of class imbalance were used, calculated using the 'data_distribution' column in our dataset. The outputs of these equations were then summed to create a class imbalance index (*i*) for each case study.

3. RESULTS

3.1. TEMPORAL AND METHODOLOGICAL TRENDS

The period from the initial application of ML in use-wear and BSM analysis to its more systematic and continuous integration into the field spanned twenty years, from 1998 to 2018 (Figure 1). During this period, ML publications ranged from zero to two papers annually. Between 2018 and 2020, the publication rate of ML papers increased rapidly, reaching up to eight publications per year, followed by a slight decrease to five papers annually from 2021 to 2022. Between 2022 and 2024, the number of publications again showed an upward trend, reaching nine papers annually.

3.1.1. Machine learning and open science approaches

In terms of the types of ML algorithms used (see Table 1), from 1998 to 2010, researchers primarily employed basic methods such as clustering algorithms and decision trees, alongside the earliest case studies using NN (Figure 2). In 2018, a noticeable increase and diversification in the application of ML methods were observed, without a clear preference for any specific approach. This balance shifted from 2019 onward, with a notable increase in both the

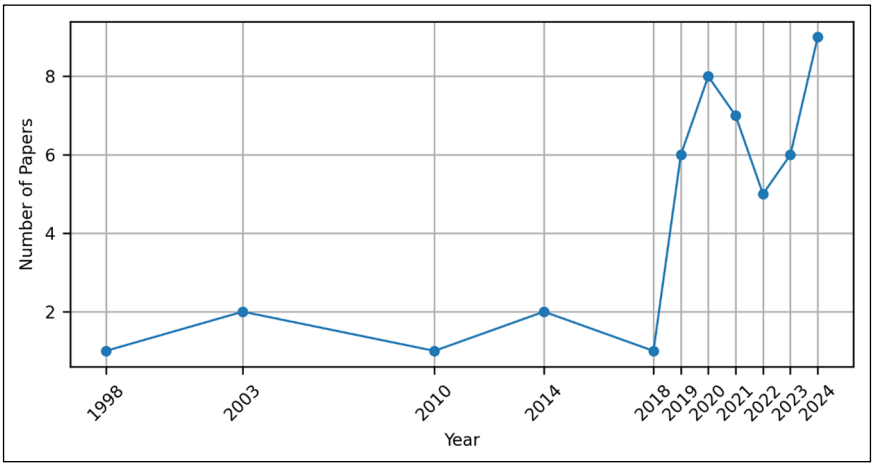

Figure 1 Annual publication rate of use-wear and BSM papers using machine learning.

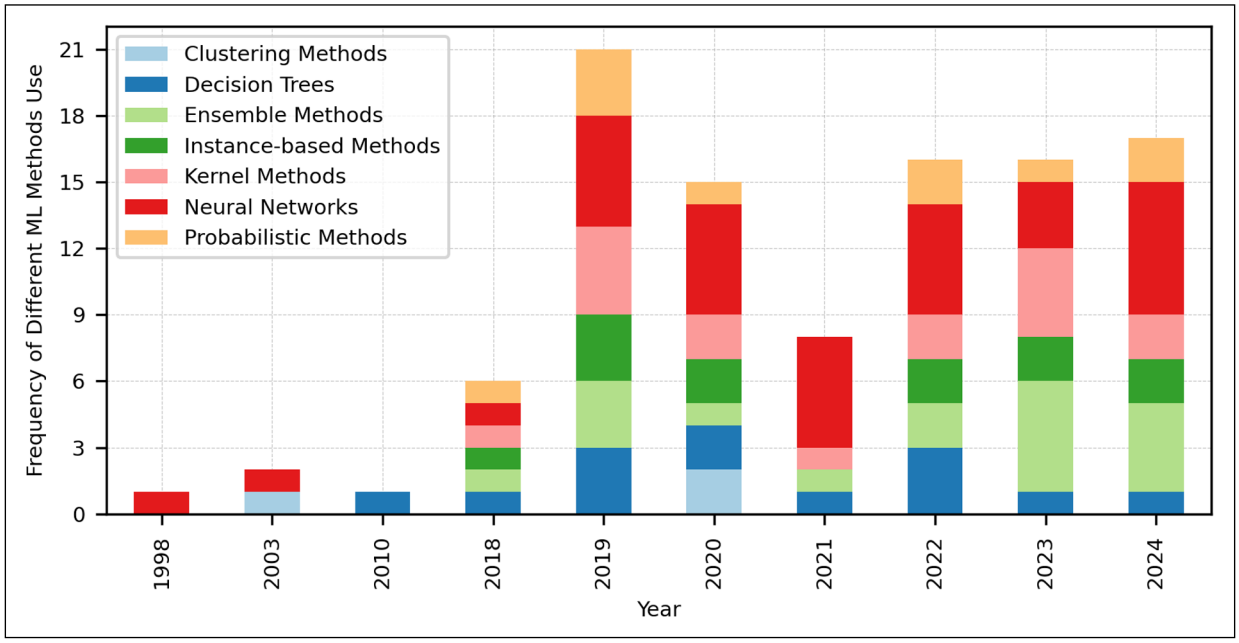

Figure 2 Temporal trends in the frequency and categories of machine learning algorithms used in use-wear and BSM papers.

frequency and variety of ML algorithms used, including NN, kernel-based methods, ensemble algorithms, and instance-based approaches. The only method showing a nearly consistent preference was NN, which appears to dominate the landscape of use-wear and BSM papers, with 10 articles employing them in 2024. There are two main trends regarding the analytical tools used based on their type and accessibility.

Among the 48 papers reviewed in this study, 38 used a programming language for their analysis, 9 used software, and 1 paper did not specify the method (Figure 3). Open science practices were found to be limited, with 16 out of 38 papers using a programming language for analysis failing to provide their scripts, and 7 out of 9 software-based papers using licensed software.

In terms of the preferred programming languages, MATLAB and Prolog were used in one paper each, R was used in 19 papers, and Python in 24 papers (Figure 4). The choice of programming language appears to be related to the dimensionality of the input data, with Python being preferred primarily for papers involving 2D and 3D imagery, while R is preferred for tabular data.

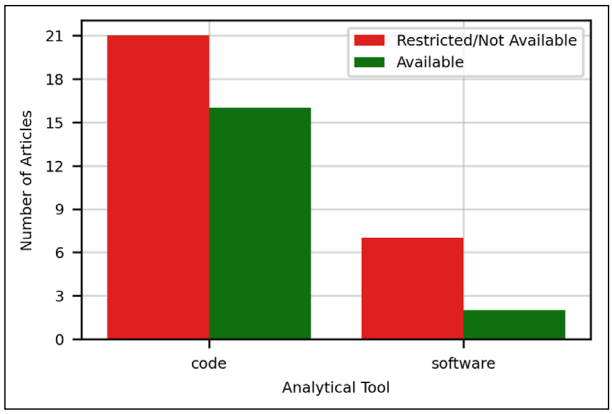

Figure 3 Number of papers using software or programming languages for analysis, categorized by whether the software is open source or licensed, and whether the code is shared openly or not.

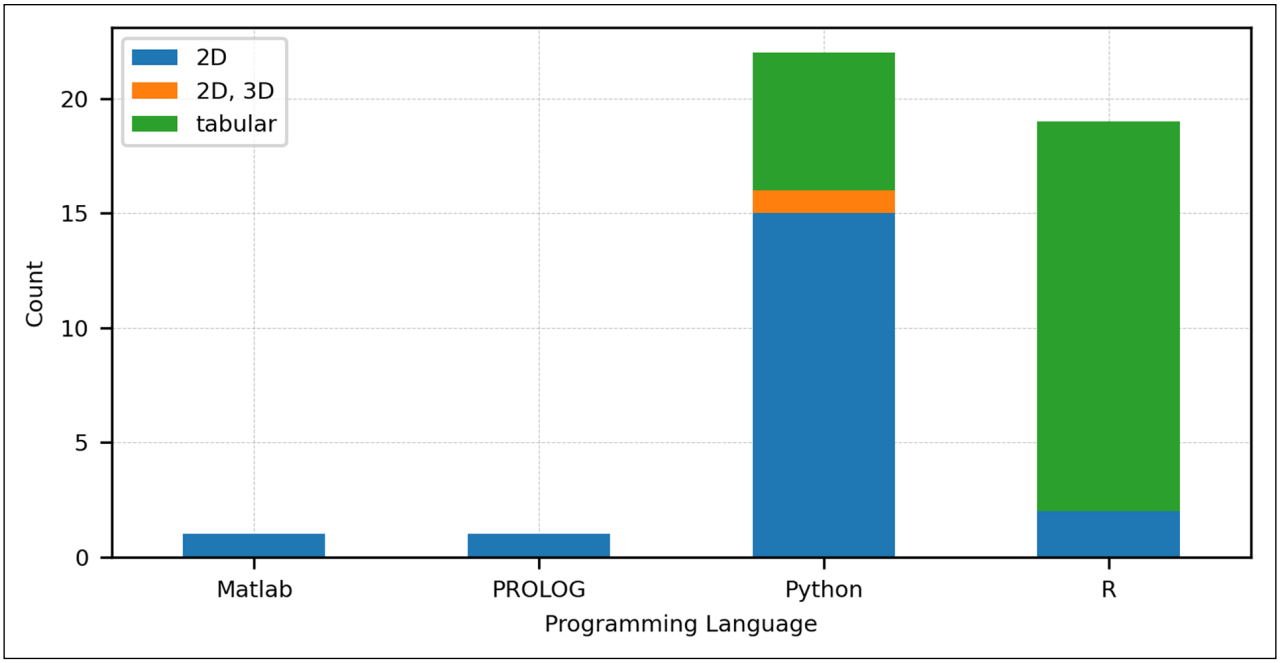

Figure 4 Programming languages used in the papers under review, grouped by data type (tabular, 2D images, or 3D images).

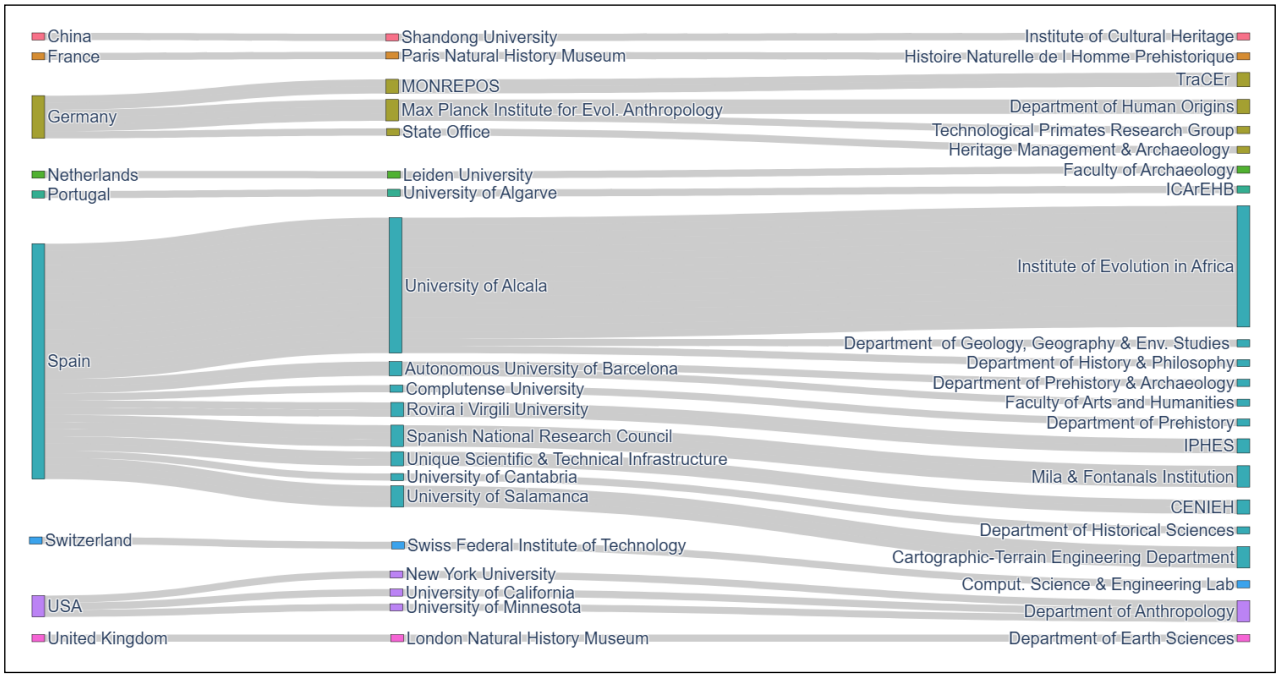

Figure 5 Frequency of countries, institutions, and laboratories or departments involved in publishing use-wear and BSM articles utilizing machine learning.

3.1.2. Geographical and institutional distribution of publications

The data exhibited recurring patterns in the institutions contributing to the field, the research topics explored, and the algorithms employed. Spain leads in the publication rate of articles, with the University of Alcalá accounting for nearly half of these publications, followed by Germany, and the United States (Figure 5). A smaller proportion of the total publications came from China, France, the Netherlands, Portugal, Switzerland, and the United Kingdom.

A preference was also observed in the materials being studied, with 33 papers focused on bone materials, 14 papers on lithics, and 1 paper dealing with wood. The majority of articles on bones focus on tooth marks, cut marks, and post-depositional effects (Figure 6), which are mostly explored with NN, followed by support vector machines (SVMs), and decision trees (DT) (Figure 7). The articles on lithic materials display a narrower diversity, mainly exploring the use-wear of polish on flint using DT and NN.

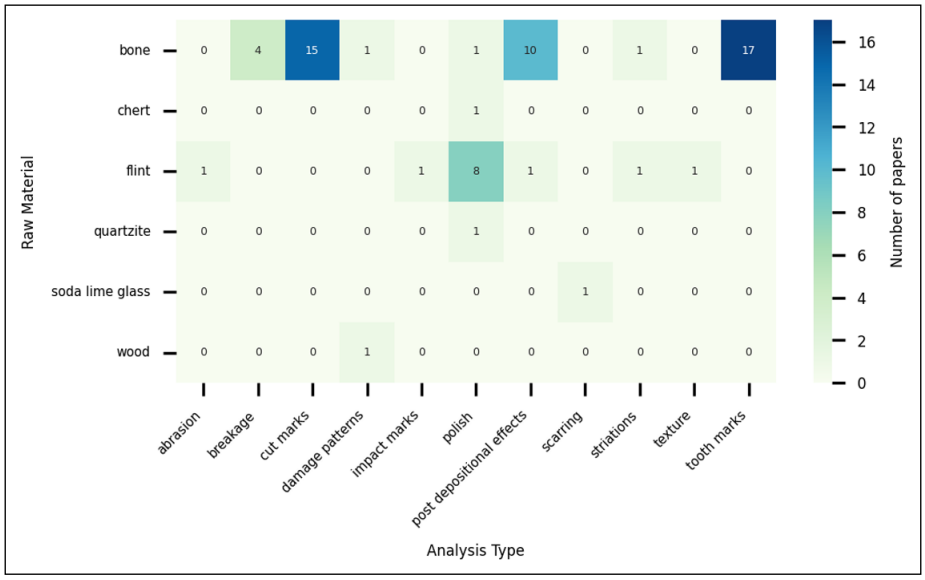

Figure 6 Heatmap displaying the frequency of research topics explored across different raw materials.

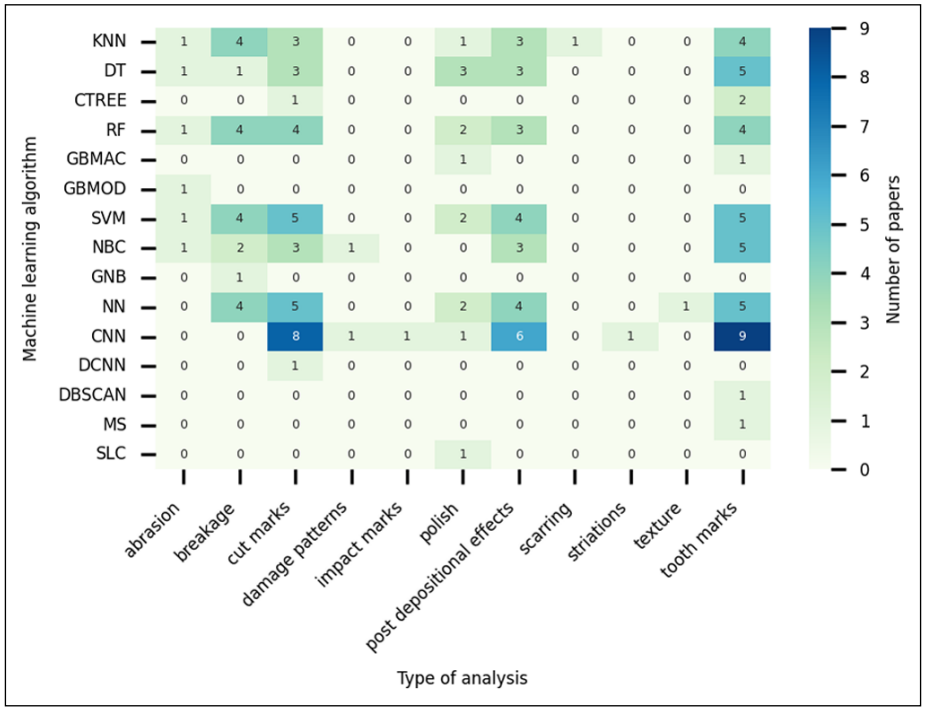

Figure 7 Heatmap illustrating the frequency of different machine learning algorithms applied to various research questions.

3.1.3. Limitations and challenges

The authors of the reviewed papers indicated that the most common limitations they encountered were related to the lack of open science practices and methodological issues, such as the use of small and unbalanced datasets (Figure 8).

Other issues affecting the application of ML in use-wear analysis and BSM pertain to archaeological constraints, including equifinality, taphonomy, and the palimpsest phenomenon. Five out of fourteen limitations were related to the data acquisition process, including the type of raw material used, absence of protocol, user bias, and limitations of the analysis or acquisition mode. A small percentage is attributed to the resources needed for the development of ML models, particularly the costs associated with time and required training.

3.2. ANALYSIS OF CLASS IMBALANCE AND DATASET SIZE

Figure 9 illustrates the relationship between model accuracy, class imbalance, and dataset size in the reviewed papers. Across all studies, accuracy ranged from 65% to 100%, and class imbalance from 1 to 15, with no apparent trend linking dataset size to accuracy or class imbalance. Most studies utilized datasets ranging from 60 to 500 samples, with reported accuracies spanning from 40% to 100%, though the majority exceeded 70%. Class imbalance varied between values of 1 and 6.

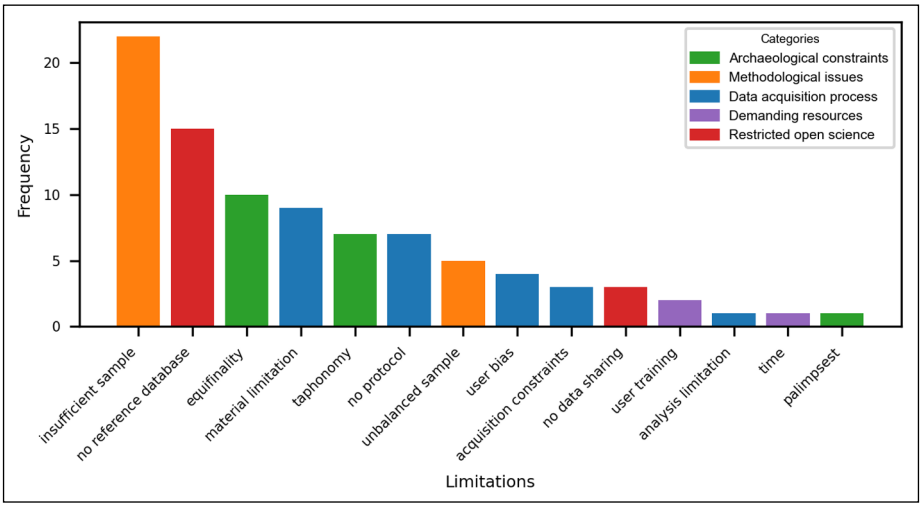

Figure 8 Frequency of distinct types of limitations discussed in the literature regarding the effective application of machine learning in use-wear and BSM analysis.

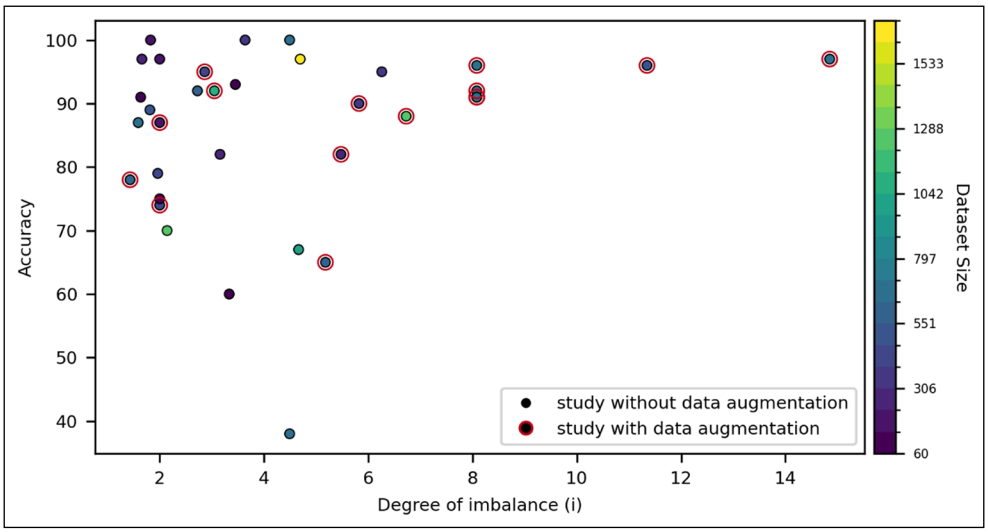

Figure 9 Scatterplot illustrating the relationship between class imbalance, dataset size, and the accuracy achieved in each case study.

However, five studies exhibited higher imbalance ratios between 7 and 15, a range not observed in studies with larger datasets. Four studies used datasets exceeding 1000 samples, with only one surpassing 1500. These medium-to-large datasets reported accuracy values between 69% and 96%, with class imbalance ratios ranging from 2 to 7. Data augmentation, indicated by red-outlined points, was predominantly applied in studies with smaller datasets, with only two studies employing it for medium-sized datasets.

3.3. QUALITY ASSESSMENT

The type and number of evaluation metrics used in the case studies varied. The majority of the papers (41 of 48) included at least one evaluation metric, with the most commonly occurring being accuracy, F1 score, and sensitivity (Figure 10).

In terms of the number of different metrics used, 11 papers included five or more metrics in their analysis, 11 papers included four metrics, eight papers included three metrics, and 11 papers included two or fewer metrics. Additionally, 12 out of 41 papers included learning curve plots in their publication, suggesting that the metrics in the majority of papers were reported only within the text without detailed documentation on how the values were derived during model training and validation. Consequently, our evaluation of the model performance was based solely on publications that included relevant plots.

As illustrated in Figure 11, almost half of the case studies, represented in red and orange, exhibited low *r* (roughness) values ranging from 0 to 0.2 (0 is rough and 1 is smooth), with one case study showing an *r* value of 0.39.

In contrast, the remaining four case studies, indicated in cream and green, demonstrate medium to high *r* values, specifically ranging from 0.4 to 0.6 and from 0.7 to 1, respectively. Figure 12 provides valuable insights into the performance of the loss curves, illustrating both the final loss values achieved by the models and the gap between the training and validation curves.

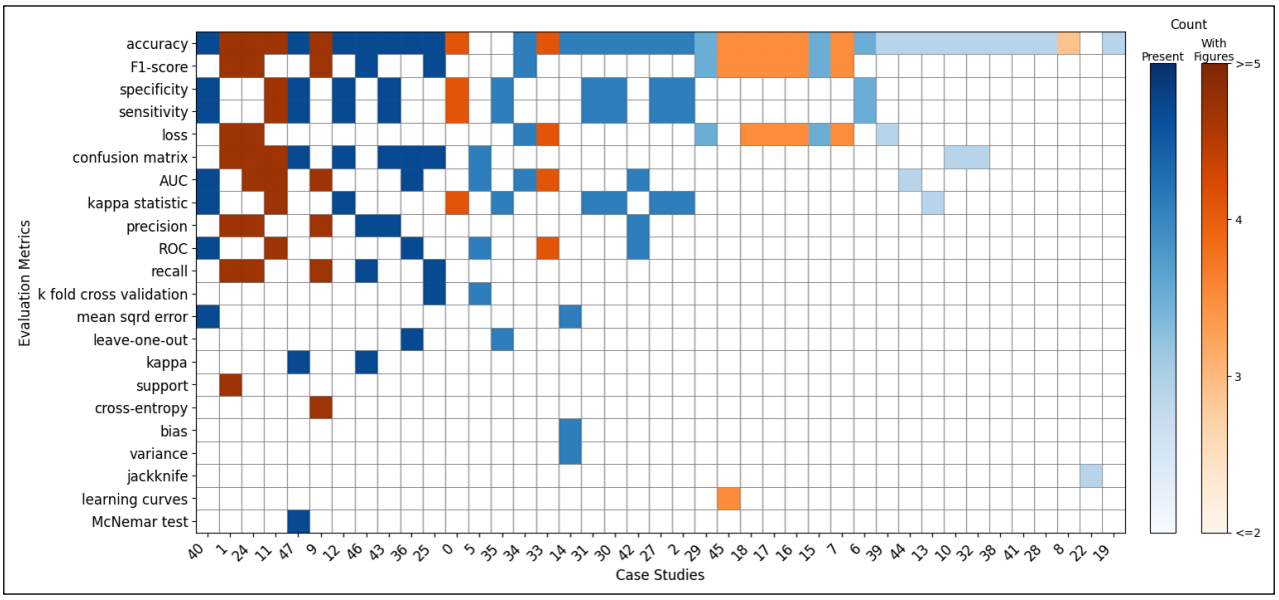

Figure 10 Matrix heatmap on the use and documentation of evaluation metrics. Papers using the largest number of metrics appear on the right, shaded in darker colors, while papers with fewer metrics (down to only one metric) are on the left with lighter shades. Blue squares indicate the presence of a metric, and orange squares indicate the presence of a metric along with supporting graphs rather than solely numerical values.

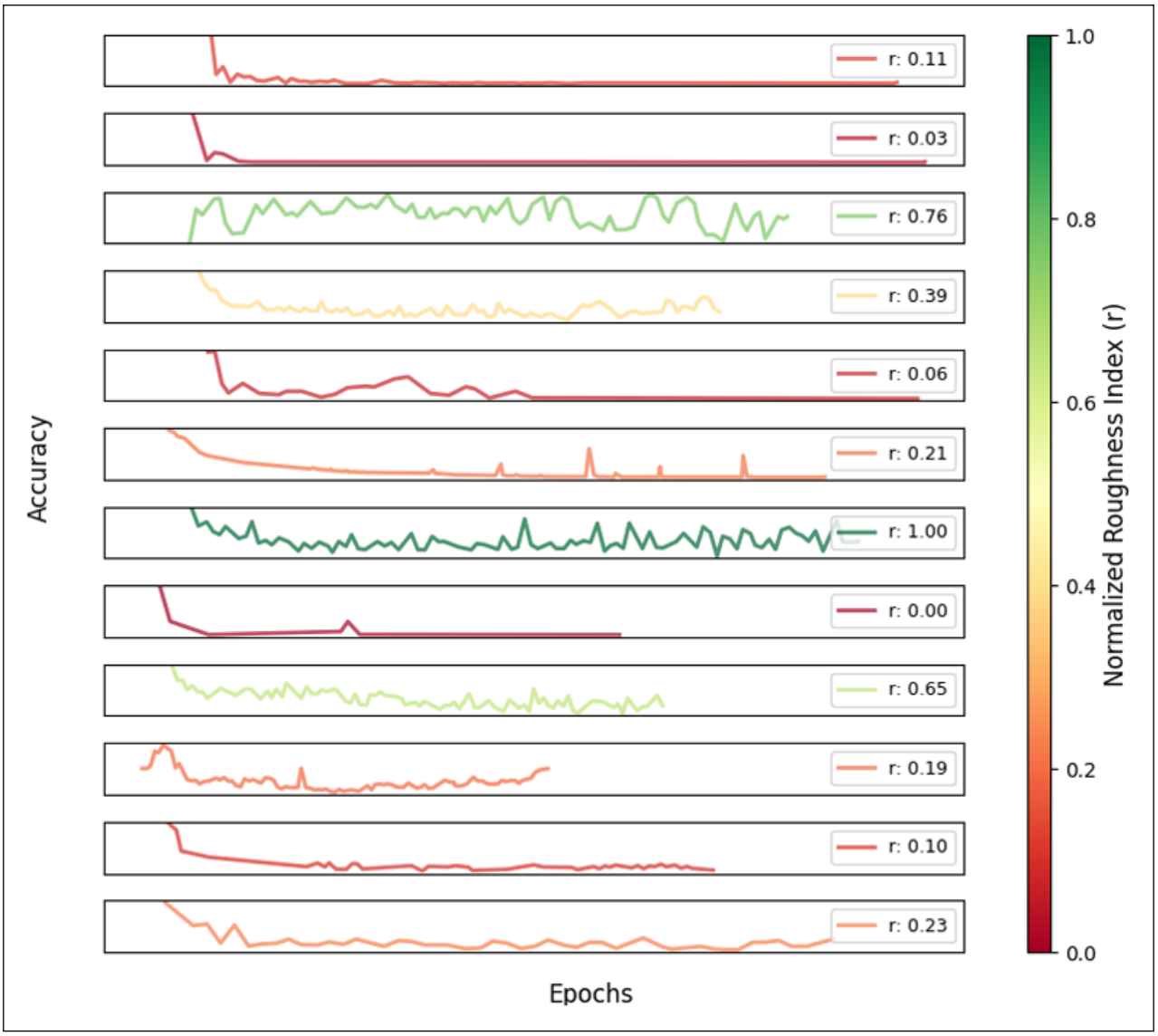

Figure 11 Roughness Index (*r*) calculated for the validation loss curves present in the reviewed papers. Smoother curves, indicated in red and orange, have lower r values. More irregular, rough curves, shown in yellow and green, have r values closer to 1.

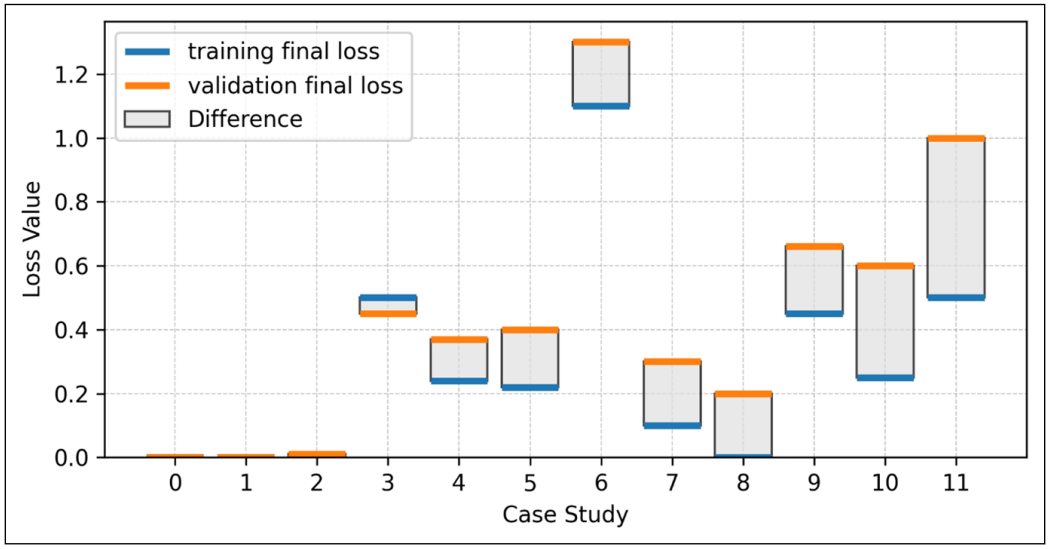

Figure 12 Boxplot illustrating the final loss values mentioned in the publications for both the training (blue line) and validation curves (orange line) at the end of each model run.

Eight out of ten case-studies show a small gap between their loss curves ranging between 0 and 0.2 degrees and two cases showing a larger gap between 0.4 and 0.8 degrees. In terms of the final loss values achieved by the models, there are four case studies where their value is below 0.2. The other six case studies demonstrated relatively low loss values ranging between 0.2 and 0.7, and only two cases exceeded 1.0.

4. DISCUSSION

This paper aims to explore the state-of-the-art in ML for use-wear analysis and BSM along two key dimensions: analyzing its initial introduction, development and application over time, and assessing its application. The introduction of ML in use-wear analysis exhibits similarities, but also some differences, with the trajectory of ML in archaeology in general. Argyrou & Agapiou (2022) created a timeline showing the annual publications of ML in archaeology from 2000 to 2020 based on two databases from Scopus of Elsevier (https://www.elsevier.com/products/scopus) and Dimensions project (https://www.dimensions.ai/; see Supplementary Materials Figures 1 and 2). The overlap in the timeframes used in our review (from 1998 to 2024) and by Argyrou & Agapiou (2022) (from 2000 to 2022) offers an opportunity to examine how our work fits the wider discussion of ML in archaeology. Scopus' results are comparable with Figure 1, showing the highest publication rate in 2019. However, the publication rate in archaeology is increasing at a smoother pace, peaking in 2020 with almost double the number of papers than the year before. The case of use-wear analysis and BSM differs, exhibiting a plateau in the publication of ML papers from 1998 to 2018 with very few fluctuations until it reached a peak in 2020 with eight papers, compared to six papers in 2019 and one paper in 2018. The results from the Dimensions project show a more gradual trajectory in the publication of ML papers in archaeology, with a minimum of 400 papers already from 2001 and with no apparent abrupt increase. The number of publications included in our study and that of Argyrou & Agapiou (2022) could be expanded by incorporating non-English sources, which may be considered in future research.

The body of ML publications in use-wear analysis and BSM is relatively small and uneven in terms of materials, algorithms, research questions, and contributing institutions. Current studies have mainly focused on flint and polish wear. Expanding research to include a wider range and greater quantity of raw materials (e.g., quartz, obsidian) and use-wear types (e.g., striations and edge damage) would better align with the diversity observed in the archaeological record and enhance model robustness by exposing algorithms to a broader spectrum of parameters, thereby improving their ability to generalize across different contexts (Rodriguez et al. 2020). Bearing in mind that Wolpert's 'no free lunch' theorem (Wolpert 1996) states that there is no universally best model in ML that performs effectively on every task, a more comprehensive dataset would facilitate a systematic evaluation of ML performance across different archaeological materials, leading to more effective and context-specific applications.

The potential of ML in use-wear studies is further constrained by its focus mainly on classification tasks. Over half of the ML papers (25 out of 48) in bone, wood and lithic materials use datasets of traces identified manually by researchers, a process that is time-consuming and prone to inter- and intra-user bias (Abellán et al. 2021; Calandra et al. 2019b; Courtenay et al. 2020a, 2020b; Marreiros, Pereira & Iovita 2020). Researchers in other subfields of archaeology, predominately in remote sensing (Guyot et al. 2021) but also in geoarchaeology

(Zickel et al. 2024), archaeobotany (Barron 2023) and osteoarchaeology (Tanti et al. 2021), are using image segmentation in order to automate the process of identifying features which can be either macroscopic like structures, mounds, skeletal members or microscopic such as minerals or anatomical parts of grains. Image segmentation involves partitioning an image into separate regions (i.e., segments) by identifying individual features or boundaries that can later be measured or classified (Merchant & Castleman 2022). Use-wear analysts could thus benefit from expanding the use of ML in multiple aspects of their work in order to limit user involvement and increase efficiency (Zhang et al. 2024). Another significant step towards enhancing ML in use-wear analysis could involve addressing the concentration of publications within a limited academic community (Argyrou & Agapiou 2022). Increasing multivocality in research by encouraging contributions from both more and non-solely Western institutions will enrich the quantity and quality of research by offering a larger and more diverse community to engage in discussion, thus advancing the field.

A major issue highlighted in the review is the current use of machine learning in use-wear analysis within the framework of open science, which presents two key challenges. First, the limited availability of large and diverse datasets contributes to class imbalance and overfitting. Open science practices can help mitigate these issues by depositing datasets in publicly accessible repositories using standardized formats (e.g., .csv, .json) and including detailed metadata (Calandra et al. 2019a). Establishing a dedicated open-access repository for use-wear and bone surface modification (BSM) data, similar to IsoBank for isotopes (Shipley et al. 2024), the European Pollen Database for pollen (Davis et al. 2013) and Neotoma for palaeoecological data (Goring 2018), could further facilitate data sharing (Calandra et al. 2019a). While challenges such as data standardization, storage costs, and long-term sustainability exist, these can be addressed through strategic planning, standardized workflows and ontologies, and sustainable funding models led by individual labs or dedicated projects (Kintigh 2006; Nicholson et al. 2023).

Second, there is an evident preference for using proprietary software (e.g., SPSS, MATLAB) and limited sharing of coding scripts. While proprietary software has some advantages, such as user-friendly interfaces and technical support, adopting open-source alternatives (e.g., Orange Data mining), where feasible, could improve transparency and reproducibility (Amandeep, Bansal & Neetu 2015; Marwick 2017). Prioritizing script-based, openly published workflows would further strengthen these outcomes by providing a detailed and verifiable record of the workflow followed (Marwick 2017). The advancement of open science practices can be further supported through the involvement of institutions, academic journals, and funding bodies. Specifically, structured training programs (e.g., modules from the Center for Open Science; https://www.cos.io/)), guidelines that encourage open-access publishing and data sharing (such as those implemented by the Journal of Computer Applications in Archaeology), and recognition systems that reward transparent research practices (e.g., the Journal of Archaeological Science Reproducibility Prize) can contribute to these efforts (Marwick et al. 2022).

The use of evaluation metrics in the papers under review shows a positive development, with some studies utilizing a variety of metrics that are thoroughly documented in their publications. However, it is important to note that a significant proportion of the studies relied on a small number of metrics and/or lacked comprehensive documentation. The limited use of evaluation metrics, in particular, can raise concerns, as 'if a study presents a single metric, one might question the performance of the classifier when evaluated using other metrics' (Lever, Krzywinski & Altman 2016: 603). Limited access to information regarding a model's performance during training and evaluation restricts our ability to fully assess its effectiveness (Viering & Loog 2023). Therefore, it is essential to employ a diverse set of evaluation metrics that align with the needs of our data and are presented in publications through graphs or tables (Lever, Krzywinski & Altman 2016).

Overall, the papers under review present models of acceptable robustness and quality, capable of meeting their goals in classifying use-wear traces and BSM. Nevertheless, some models exhibit instability, as evidenced by irregular learning curves and gaps between training and validation loss curves, which can be connected to the use of unsuitable learning rates or small datasets (Goodfellow, Bengio & Courville 2016). The occurrence of gaps between training and validation curves may suggest potential overfitting where the model is memorizing the dataset, including its noise (Singh 2019). Furthermore, the high accuracies reported in some studies should be approached with caution, as they may arise from the use of small, unrepresentative or unbalanced datasets and multicollinearity (correlated samples; Ellis, Sander & Limon 2022; Goodfellow, Bengio & Courville 2016; Walsh, Pollastri & Tosatto 2016). Future research could benefit from exploring additional dimensions of ML model performance, such as underfitting, interpretability, and misclassification patterns, along with the influence of factors, such as the type of raw material, use-wear trace, and acquisition methods. The identification and resolution of these issues can be tackled through interdisciplinary collaboration, as partnerships with computer scientists and computational archaeologists can facilitate the integration of relevant scientific expertise (Courtenay et al. 2020a). Because each dataset is unique and influenced by its acquisition methods and material characteristics

(Lever, Krzywinski & Altman 2016), some studies may benefit from collaborations with researchers who can tailor the design, implementation, and interpretation of the ML workflow to address the specific biases and requirements of the data in question (Courtenay et al. 2020a). By doing so, we will be able to effectively integrate ML into archaeological practice, ensuring that our analyses are thoughtfully aligned with the archaeological questions at hand, rather than being 'determined by what the technology will do' (Lock 2009: 82).

The efficiency of a ML model relies not only on its correct design and tuning but also on the input of high-quality data (Merchant & Castleman 2022). Aspects such as pixel spacing, resolution, noise, photometry, and distortion affect the overall quality of images captured by microscopes (ibid.) and can directly impact a model's ability to identify or classify various traces (Bustos-Pérez & Ollé 2024; Courtenay et al. 2020a). The systematic acquisition of experimental and analytical data is vital to ensure their quality and reproducibility (Marreiros et al. 2020) as well as to provide homogeneous data for building ML models. Failure to do so may lead to incorrect or biased results, emphasizing the necessity for use-wear analysts to adhere to thorough documentation and acquisition practices.

5. CONCLUSION

This article presents a review of ML applications in use-wear analysis of lithic, bone and wood materials, aiming to provide constructive suggestions for expanding and strengthening its application. To achieve a comprehensive understanding of the state-of-the-art, we examined the origins and evolution of ML use in use-wear analysis and BSM studies. Using standard descriptive statistics, notable patterns were detected, highlighting the limited adoption of open science practices, the concentration of the publishing community around a few, geographically localized institutions, and the focus on specific research topics and materials.

Our contribution marks a notable advancement in the field of computational archaeology by (a) presenting a comprehensive review of ML applications specifically in use-wear analysis and BSM, and (b) aiding archaeologists in understanding the key components of a ML workflow, as well as the critical factors essential for successfully integrating these methods into their research. A set of quantitative approaches were employed to assess and compare the robustness of models from different research papers, enabling the identification of potential issues and areas for improvement. The absence of large and diverse datasets significantly affects the quality of the models being developed and the research questions being addressed, highlighting the importance of openly sharing data and ensuring their quality. Additionally, collaboration and interdisciplinarity emerged as essential factors for improving the efficiency of the models, given the need for specialized skills and understanding of ML concepts and workflows.

To fully benefit from ML's ability to reduce bias, automate tasks, and perform powerful analyses, it is crucial to acknowledge and understand its limitations and requirements. This study provides a foundation for informed decision making regarding the use of ML in archaeology, covering important issues and aspects from data acquisition to model development and assessment. Since ML applications typically require more data than individual archaeological projects can generate (Argyrou & Agapiou 2022), collaboration through open science is not merely beneficial but essential for developing robust models (McPherron et al. 2022). Expanding and diversifying data repositories could enhance ML model performance (Courtenay et al. 2020a), while also strengthening research transparency, reproducibility, and accessibility (Marwick 2017). Developing reliable models can facilitate the study of archaeological materials by enabling the effective handling of larger datasets, automating the identification of features and patterns in a standardized manner, and allowing researchers to explore more complex research questions and ideas. Achieving this goal requires open-mindedness, communication, and transparency to enable ML to unfold its full potential in use-wear analysis and archaeology.

DATA ACCESSIBILITY STATEMENT

The Quarto file and associated data used in the creation of this paper are openly available in a Zenodo repository (DOI: https://zenodo.org/records/15458673).

ADDITIONAL FILE

The additional file for this article can be found as follows:

- **Supplementary material.** Figures S1 and S2. DOI: https://doi.org/10.5334/jcaa.190.s1

FUNDING INFORMATION

This research was funded in whole by the Fundação para a Ciência e a Tecnologia, I.P. (FCT, https://ror.org/00snfqn58) under Grant UIDP/04211/2020, through the Interdisciplinary Center for Archaeology and the Evolution of Human Behaviour (ICArEHB). Open Access to this article is financed by FCT – Fundação para a Ciência e a Tecnologia, within the scope of the project UID/04211: Centro Interdisciplinar de Arqueologia e Evolução do Comportamento Humano (ICArEHB).

COMPETING INTERESTS

The authors have no competing interests to declare.

AUTHOR CONTRIBUTIONS

AE: conceptualization, methodology, formal analysis, visualization, writing the original draft, reviewing and editing the manuscript. SPM: supervision, conceptualization, methodology, reviewing and editing the manuscript. JM: supervision, conceptualization, methodology, reviewing and editing the manuscript.

AUTHOR AFFILIATIONS

Anastasia Eleftheriadou orcid.org/0000-0002-8649-3752
ICArEHB, University of Algarve, PT

Shannon P. McPherron orcid.org/0000-0002-2063-468X
Department of Human Origins, Max Planck Institute for Evolutionary Anthropology, Deutscher Platz 6, Leipzig 04103, DE

João Marreiros orcid.org/0000-0002-3399-8765
TraCEr, Laboratory for Traceology and Controlled Experiments, MONREPOS, D-56567 Neuwied, DE; Archaeological Research Centre, and Museum for Human Behavioural Evolution – LEIZA, Schloss Monrepos, D-56567 Neuwied, DE

REFERENCES

Abellán, N, Jiménez-García, B, Aznarte, J, Baquedano, E and **Domínguez-Rodrigo, M.** 2021. 'Deep learning classification of tooth scores made by different carnivores: achieving high accuracy when comparing African carnivore taxa and testing the hominin shift in the balance of power'. *Archaeological and Anthropological Sciences*, 13(2): 31. DOI: https://doi.org/10.1007/s12520-021-01273-9

Amandeep, S, Bansal, RK and **Neetu, J.** 2015. 'Open source software vs proprietary software'. *International Journal of Computer Applications*, 114(18): 26–31. DOI: https://doi.org/10.5120/20080-2132

Ambrose, SH. 2001. 'Paleolithic Technology and Human Evolution'. *Science*, 291(5509): 1748–1753. DOI: https://doi.org/10.1126/science.1059487

Argyrou, A and **Agapiou, A.** 2022. 'A review of artificial intelligence and remote sensing for archaeological research'. *Remote Sensing*, 14(23): 6000. DOI: https://doi.org/10.3390/rs14236000

Bankhead, P, Loughrey, MB, Fernández, JA, Dombrowski, Y, McArt, DG, Dunne, PD, McQuaid, S, Gray, RT, Murray, LJ, Coleman, HG, James, JA, Salto-Tellez, M and **Hamilton, PW.** 2017. 'QuPath: Open source software for digital pathology image analysis'. *Scientific Reports*, 7(1): 16878. DOI: https://doi.org/10.1038/s41598-017-17204-5

Barron, A. 2023. 'Applications of microct imaging to archaeobotanical research'. *Journal of Archaeological Method and Theory*, 316: 557–592. DOI: https://doi.org/10.1007/s10816-023-09610-z

Berganzo-Besga, I, Orengo, H, Lumbreras, F, Carrero-Pazos, M, Fonte, J and **Vilas-Estévez, B.** 2021. 'Hybrid MSRM-Based Deep Learning and Multitemporal Sentinel 2-Based Machine Learning Algorithm Detects Near 10k Archaeological Tumuli in North-Western Iberia'. *Remote Sensing*, 13(20): 4181. DOI: https://doi.org/10.3390/rs13204181

Bini, SA. 2018. 'Artificial intelligence, machine learning, deep learning, and cognitive computing: What do these terms mean and how will they impact health care?' *The Journal of Arthroplasty*, 33(8): 2358–2361. DOI: https://doi.org/10.1016/j.arth.2018.02.067

Bishop, CM. 2006. *Pattern recognition and machine learning*. New York: Springer.

Bonney, H. 2014. 'An investigation of the use of discriminant analysis for the classification of blade edge type from cut marks made by metal and bamboo blades'. *American Journal of Physical Anthropology*, 154(4): 575–584. DOI: https://doi.org/10.1002/ajpa.22558

Borel, A, Deltombe, R, Moreau, P, Ingicco, T, Bigerelle, M and **Marteau, J.** 2021. 'Optimization of use-wear detection and characterization on stone tool surfaces'. *Scientific Reports*, 11(1): 24197. DOI: https://doi.org/10.1038/s41598-021-03663-4

Buda, M, Maki, A and **Mazurowski, MA.** 2018. 'A systematic study of the class imbalance problem in convolutional neural networks'. *Neural Networks*,106: 249–259. DOI: https://doi.org/10.1016/j.neunet.2018.07.011

Bustos-Pérez, G and **Ollé, A.** 2024. 'The quantification of surface abrasion on flint stone tools'. *archaeometry*, 66(2): 247–265. DOI: https://doi.org/10.1111/arcm.12913

Bzdok, D, Altman, N and **Krzywinski, M.** 2018. 'Statistics versus machine learning'. *Nature Methods*, 15(4): 233–234. DOI: https://doi.org/10.1038/nmeth.4642

Calandra, I, Schunk, L, Bob, K, Gneisinger, W, Pedergnana, A, Paixao, E, Hildebrandt, A and **Marreiros, J.** 2019a. 'The effect of numerical aperture on quantitative use-wear studies and its implication on reproducibility'. *Scientific Reports*, 9(1): 1–10. DOI: https://doi.org/10.1038/s41598-019-42713-w

Calandra, I, Schunk, L, Rodriguez, A, Gneisinger, W, Pedergnana, A, Paixao, E, Pereira, T, Iovita, R and **Marreiros, J.** 2019b. 'Back to the edge: relative coordinate system for use-wear analysis'. *Archaeological and Anthropological Sciences*, 11: 5937–5948. DOI: https://doi.org/10.1007/s12520-019-00801-y

Calder, J, Coil, R, Melton, JA, Olver, PJ, Tostevin, G and **Yezzi-Woodley, K.** 2022. 'Use and misuse of machine learning in anthropology'. *IEEE BITS the Information Theory Magazine*, 2(1): 102–115. DOI: https://doi.org/10.1109/MBITS.2022.3205143

Caneva, I. 2001. *Beyond tools: redefining the PPN lithic assemblages of the Levant : proceedings of the Third Workshop on PPN Chipped Lithic Industries, Department of Classical and Near Eastern Studies, Ca' Foscari University of Venice, 1st – 4th November, 1998*. Berlin: Ex Oriente.

Caruso Fermé, L and **Aschero, C.** 2020. 'Manufacturing and use of the wooden artifacts. A use-wear analysis of wood technology in hunter-gatherer groups (Cerro Casa de Piedra 7 site, Argentina)'. *Journal of Archaeological Science: Reports*, 31: 102291. DOI: https://doi.org/10.1016/j.jasrep.2020.102291

Cole, V and **Boutet, M.** 2023. 'ResearchRabbit (product review)'. *The Journal of the Canadian Health Libraries Association*, 44(2): 43–47. DOI: https://doi.org/10.29173/jchla29699

Courtenay, LA, Herranz-Rodrigo, D, Huguet, R, Maté-González, MÁ, González-Aguilera, D and **Yravedra, J.** 2020a. 'Obtaining new resolutions in carnivore tooth pit morphological analyses: A methodological update for digital taphonomy'. *PloS one*, 15(10): e0240328. DOI: https://doi.org/10.1371/journal.pone.0240328

Courtenay, LA, Huguet, R, González-Aguilera, D and **Yravedra, J.** 2020b. 'A hybrid geometric morphometric deep learning approach for cut and trampling mark classification'. *Applied Sciences*, 10(1): 150. DOI: https://doi.org/10.3390/app10010150

Courtenay, LA, Yravedra, J, Huguet, R, Aramendi, J, Maté-González, MÁ, González-Aguilera, D and **Arriaza, MC.** 2019. 'Combining machine learning algorithms and geometric morphometrics: A study of carnivore tooth marks'. *Palaeogeography, Palaeoclimatology, Palaeoecology*, 522: 28–39. DOI: https://doi.org/10.1016/j.palaeo.2019.03.007

Davis, BAS, Zanon, M, Collins, P, Mauri, A, Bakker, J, Barboni, D, Barthelmes, A, Beaudouin, C, Bjune, AE, Bozilova, E, Bradshaw, RHW, Brayshay, BA, Brewer, S, Brugiapaglia, E, Bunting, J, Connor, SE, de Beaulieu, J-L, Edwards, K, Ejarque, A, Fall, P, Florenzano, A, Fyfe, R, Galop, D, Giardini, M, Giesecke, T, Grant, MJ, Guiot, J, Jahns, S, Jankovská, V, Juggins, S, Kahrmann, M, Karpińska-Kołaczek, M, Kołaczek, P, Kühl, N, Kuneš, P, Lapteva, EG, Leroy, SAG, Leydet, M, Guiot, J, López Sáez, JA, Masi, A, Matthias, I, Mazier, F, Meltsov, V, Mercuri, AM, Miras, Y, Mitchell, FJG, Morris, JL, Naughton, F, Nielsen, AB, Novenko, E, Odgaard, B, Ortu, E, Overballe-Petersen, MV, Pardoe, HS, Peglar, SM, Pidek, IA, Sadori, L, Seppä, H, Severova, E, Shaw, H, Święta-Musznicka, J, Theuerkauf, M, Tonkov, S, Veski, S, van der Knaap, WO, van Leeuwen, JFN, Woodbridge, J, Zimny, M and **Kaplan, JO.** 2013. 'European modern pollen database (EMPD) project'. *Vegetation History and Archaeobotany*, 22(6): 521–530. DOI: https://doi.org/10.1007/s00334-012-0388-5

Ellis, RJ, Sander, RM and **Limon, A.** 2022. 'Twelve key challenges in medical machine learning and solutions'. *Intelligence-Based Medicine*, 6: 100068. DOI: https://doi.org/10.1016/j.ibmed.2022.100068

Evans, AA. 2014. 'On the importance of blind testing in archaeological science: the example from lithic functional studies'. *Journal of Archaeological Science*, 48: 5–14. DOI: https://doi.org/10.1016/j.jas.2013.10.026

Evans, AA and **Donahue, RE.** 2008. 'Laser scanning confocal microscopy: a potential technique for the study of lithic microwear'. *Journal of Archaeological Science*, 35(8): 2223–2230. DOI: https://doi.org/10.1016/j.jas.2008.02.006

Evans, AA and **MacDonald, D.** 2011. 'Using metrology in early prehistoric stone tool research: Further work and a brief instrument comparison'. *Scanning*, 33(5): 294–303. DOI: https://doi.org/10.1002/sca.20272

Faulks, NR, Kimball, LR, Hidjrati, N and **Coffey, TS.** 2011. 'Atomic force microscopy of microwear traces on Mousterian tools from Myshtylagty Lagat (Weasel Cave), Russia'. *Scanning*, 33(5): 304–315. DOI: https://doi.org/10.1002/sca.20273

Fisher, JW. 1995. 'Bone Surface Modifications in Zooarchaeology'. *Journal of Archaeological Method and Theory*, 2(1): 7–68. DOI: https://doi.org/10.1007/BF02228434

Foley, R and **Lahr, MM.** 2003. 'On stony ground: Lithic technology, human evolution, and the emergence of culture'. *Evolutionary Anthropology: Issues, News, and Reviews*, 12(3): 109–122. DOI: https://doi.org/10.1002/evan.10108

Friedrich, S, Antes, G, Behr, S, Binder, H, Brannath, W, Dumpert, F, Ickstadt, K, Kestler, HA, Lederer, J, Leitgöb, H, Pauly, M, Steland, A, Wilhelm, A and **Friede, T.** 2022. 'Is there a role for statistics in artificial intelligence?' *Advances in Data Analysis and Classification*, 16(4): 823–846. DOI: https://doi.org/10.1007/s11634-021-00455-6

Fullagar, R. 2014. 'Residues and usewear'. In: Balme, J and Paterson, A (eds.), *Archaeology in Practice: a Student Guide to Archaeological Analyses*. 2nd edition. Malden: Blackwell Publishing, pp. 232–263. DOI: https://doi.org/10.1002/9781394260966.ch8

Ghosh, K, Bellinger, C, Corizzo, R, Branco, P, Krawczyk, B and **Japkowicz, N.** 2024. 'The class imbalance problem in deep learning'. *Machine Learning*, 113(7): 4845–4901. DOI: https://doi.org/10.1007/s10994-022-06268-8

González-Urquijo, JE and **Ibáñez-Estévez, JJ.** 2003. 'The quantification of use-wear polish using image analysis. First results'. *Journal of Archaeological Science*, 30(4): 481–489. DOI: https://doi.org/10.1006/jasc.2002.0855

Goodfellow, I, Bengio, Y and **Courville, A.** 2016. *Deep learning*. 1st ed. London: The MIT Press.

Goring, SJ. 2018. *The Neotoma Paleoecology Database: a research-outreach nexus*. Cambridge: Cambridge University Press. DOI: https://doi.org/10.1017/9781108681582

Guyot, A, Lennon, M, Lorho, T and **Hubert-Moy, L.** 2021. 'Combined detection and segmentation of archeological structures from LiDAR data using a deep learning approach'. *Journal of Computer Applications in Archaeology*, 4(1): 1–19. DOI: https://doi.org/10.5334/jcaa.64

Handelman, GS, Kok, HK, Chandra, RV, Razavi, AH, Huang, S, Brooks, M, Lee, MJ and **Asadi, H.** 2019. 'Peering into the black box of artificial intelligence: Evaluation metrics of machine learning methods'. *American Journal of Roentgenology*, 212(1): 38–43. DOI: https://doi.org/10.2214/AJR.18.20224

Harris, CR, Millman, KJ, van der Walt, SJ, Gommers, R, Virtanen, P, Cournapeau, D, Wieser, E, Taylor, J, Berg, S, Smith, NJ, Kern, R, Picus, M, Hoyer, S, van Kerkwijk, MH, Brett, M, Haldane, A, Del Río, JF, Wiebe, M, Peterson, P, Gérard-Marchant, P, Sheppard, K, Reddy, T, Weckesser, W, Abbasi, H, Gohlke, C and **Oliphant, TE.** 2020. 'Array programming with NumPy'. *Nature*, 585(7825): 357–362. DOI: https://doi.org/10.1038/s41586-020-2649-2

Hayden, B. 1979. *Lithic use-wear analysis*. New York: Academic Press.

Hovers, E and **Belfer-Cohen, A.** 2013. 'On Variability and Complexity: Lessons from the Levantine Middle Paleolithic Record'. *Current Anthropology*, 54(S8): S337–S357. DOI: https://doi.org/10.1086/673880

Ibáñez, JJ, González-Urquijo, JE and **Gibaja, J.** 2014. 'Discriminating wild vs domestic cereal harvesting micropolish through laser confocal microscopy'. *Journal of Archaeological Science*, 48(1): 96–103. DOI: https://doi.org/10.1016/j.jas.2013.10.012

Ibáñez, JJ, Lazuen, T and **González-Urquijo, J.** 2019. 'Identifying Experimental Tool Use Through Confocal Microscopy'. *Journal of Archaeological Method and Theory*, 26(3): 1176–1215. DOI: https://doi.org/10.1007/s10816-018-9408-9

Ibáñez, JJ and **Mazzucco, N.** 2021. 'Quantitative use-wear analysis of stone tools: Measuring how the intensity of use affects the identification of the worked material'. *PloS one*,16(9): e0257266. DOI: https://doi.org/10.1371/journal.pone.0257266

James, EC and **Thompson, JC.** 2015. 'On bad terms: Problems and solutions within zooarchaeological bone surface modification studies'. *Environmental Archaeology*, 20(1): 89–103. DOI: https://doi.org/10.1179/1749631414Y.0000000023

Jordahl, K, den Bossche, JV, Fleischmann, M, Wasserman, J, McBride, J, Gerard, J, Tratner, J, Perry, M, Badaracco, AG, Farmer, C, Hjelle, GA, Snow, AD, Cochran, M, Gillies, S, Culbertson, L, Bartos, M, Eubank, N, maxalbert, Bilogur, A, Rey, S, Ren, C, Arribas-Bel, D, Wasser, L, Wolf, LJ, Journois, M, Wilson, J, Greenhall, A, Holdgraf, C, Filipe and **Leblanc, F.** 2020. *geopandas/geopandas: v0.8.1*. DOI: https://doi.org/10.5281/zenodo.3946761

Keeley, LH. 1980. *Experimental Determination of Stone Tool Uses: A Microwear Analysis*. Chicago: University of Chicago Press.

Kintigh, K. 2006. 'The promise and challenge of archaeological data integration'. *American Antiquity*, 71(3): 567–578. DOI: https://doi.org/10.2307/40035365

Kubat, M. 2017. *An Introduction to Machine Learning*. 2nd ed. Cham: Springer. DOI: https://doi.org/10.1007/978-3-319-63913-0

Kuhn, SL. 2020. 'Moving on from Here: Suggestions for the Future of "MobilityThinking" in Studies of Paleolithic Technologies'. *Journal of Paleolithic Archaeology*, 3(4): 664–681. DOI: https://doi.org/10.1007/s41982-020-00060-7

Lemorini, C, Plummer, TW, Braun, DR, Crittenden, AN, Ditchfield, PW, Bishop, LC, Hertel, F, Oliver, JS, Marlowe, FW, Schoeninger, MJ and **Potts, R.** 2014. 'Old stones' song: use-wear experiments and analysis of the Oldowan quartz and quartzite assemblage from Kanjera South (Kenya)'. *Journal of Human Evolution*, 72: 10–25. DOI: https://doi.org/10.1016/j.jhevol.2014.03.002

Lever, J, Krzywinski, M and **Altman, N.** 2016. 'Classification evaluation'. *Nature Methods*, 13(8): 603–604. DOI: https://doi.org/10.1038/nmeth.3945

Lock, G. 2009. 'Archaeological computing then and now: theory and practice, intentions and tensions'. *Archeologia e Calcolatori*, 20: 75–84.

Luncz, LV, Braun, DR, Marreiros, J, Bamford, M, Zeng, C, Pacome, SS, Junghenn, P, Buckley, Z, Yao, X and **Carvalho, S.** 2022. 'Chimpanzee wooden tool analysis advances the identification of percussive technology'. *iScience*, 25(11): 105315. DOI: https://doi.org/10.1016/j.isci.2022.105315

Lycett, SJ. 2015. 'Cultural evolutionary approaches to artifact variation over time and space: Basis, progress, and prospects'. *Journal of Archaeological Science*, 56: 21–31. DOI: https://doi.org/10.1016/j.jas.2015.01.004

Ma, S, Doyon, L, Zhang, Y and **Li, Z.** 2023. 'Disentangling carcass processing activities and the state of worked hide from use-wear patterns on expedient bone tools. A preliminary experiment'. *Journal of Archaeological Science: Reports*, 49: 104027. DOI: https://doi.org/10.1016/j.jasrep.2023.104027

Marreiros, J, Calandra, I, Gneisinger, W, Paixão, E, Pedergnana, A and **Schunk, L.** 2020. 'Rethinking use-wear analysis and experimentation as applied to the study of past hominin tool use'. *Journal of Paleolithic Archaeology*, 3(3): 475–502. DOI: https://doi.org/10.1007/s41982-020-00058-1

Marreiros, J, Pereira, T and **Iovita, R.** 2020. 'Controlled experiments in lithic technology and function'. *Archaeological and Anthropological Sciences*, 12(6): 110. DOI: https://doi.org/10.1007/s12520-020-01059-5

Marreiros, JM, Gibaja Bao, JF and **Bicho, NF** (eds.). 2015. *Use-wear and residue analysis in archaeology*. Cham: Springer. DOI: https://doi.org/10.1007/978-3-319-08257-8

Marreiros, JM, Mazzucco, N, Gibaja, JF and **Bicho, N.** 2015. 'Macro and Micro Evidences from the Past: The State of the Art of Archeological Use-Wear Studies'. In: Marreiros, JM, Gibaja Bao, JF and Ferreira Bicho, N (eds.), *Use-Wear and Residue Analysis in Archaeology*. Cham: Springer, pp. 5–26. DOI: https://doi.org/10.1007/978-3-319-08257-8_2

Marwick, B. 2017. 'Computational reproducibility in archaeological research: Basic principles and a case study of their implementation'. *Journal of Archaeological Method and Theory*, 24(2): 424–450. DOI: https://doi.org/10.1007/s10816-015-9272-9

Marwick, B, Wang, L-Y, Goldstein, L and **Watrall, E.** 2022. 'How to align disciplinary ideals with actual practices'. In: Watrall, E and Goldstein, L (eds.), *Digital Heritage and Archaeology in Practice*. Gainesville: University Press of Florida. DOI: https://doi.org/10.5744/florida/9780813069302.003.0010

McPherron, SP, Archer, W, Otárola-Castillo, ER, Torquato, MG and **Keevil, TL.** 2022. 'Machine learning, bootstrapping, null models, and why we are still not 100% sure which bone surface modifications were made by crocodiles'. *Journal of Human Evolution*, 164: 103071. DOI: https://doi.org/10.1016/j.jhevol.2021.103071

Merchant, F and **Castleman, K.** 2022. *Microscope image processing*. Second edition. San Diego: Elsevier Science & Technology.

Nicholson, C, Kansa, S, Gupta, N and **Fernandez, R.** 2023. 'Will it ever be FAIR?: Making archaeological data findable, accessible, interoperable, and reusable'. *Advances in Archaeological Practice*, 11(1): 63–75. DOI: https://doi.org/10.1017/aap.2022.40

Pizarro-Monzo, M, Organista, E, Cobo-Sánchez, L, Baquedano, E and **Domínguez-Rodrigo, M.** 2022. 'Determining the diagenetic paths of archaeofaunal assemblages and their palaeoecology through artificial intelligence: an application to Oldowan sites from Olduvai Gorge (Tanzania)'. *Journal of Quaternary Science*, 37(3): 543–557. DOI: https://doi.org/10.1002/jqs.3385

Raschka, S and **Mirjalili, V.** 2019. *Python Machine Learning: Machine Learning and Deep Learning with Python, scikit-learn, and TensorFlow 2*, 3rd Edition. Birmingham: Packt Publishing.

Režek, Ž, Dibble, HL, McPherron, SP, Braun, DR and **Lin, SC.** 2018. 'Two million years of flaking stone and the evolutionary efficiency of stone tool technology'. *Nature Ecology & Evolution*, 2(4): 628. DOI: https://doi.org/10.1038/s41559-018-0488-4

Rodriguez, A, Pouydebat, E, Chacón, MG, Moncel, M-H, Cornette, R, Bardo, A, Chèze, L, Iovita, R and **Borel, A.** 2020. 'Right or left? Determining the hand holding the tool from use traces'. *Journal of Archaeological Science: Reports*, 31: 102316. DOI: https://doi.org/10.1016/j.jasrep.2020.102316

Samoili, S, Lopez Cobo, M, Gomez Gutierrez, E, De Prato, G, Martinez-Plumed, F and **Delipetrev, B.** 2020. *AI WATCH. Defining artificial intelligence. Towards an operational definition and taxonomy of artificial intelligence*. Luxembourg: Publication Office of the European Union. DOI: https://doi.org/10.2760/382730

Sayers, MW. 1995. 'On the calculation of International Roughness Index from longitudinal road profile'. In: *Transportation Research Record, No. 1501*. Washington, DC: National Academy of Sciences, pp. 1–12.

Sferrazza, P. 2025. 'Archaeological and experimental lithic microwear classification through 2D textural analysis and machine learning'. *Journal of Archaeological Method and Theory*, 32: 31. DOI: https://doi.org/10.1007/s10816-025-09701-z

Shea, JJ. 2017. *Stone tools in human evolution : behavioral differences among technological primates*. Cambridge: Cambridge University Press. DOI: https://doi.org/10.1017/9781316389355

Shipley, ON, Dabrowski, AJ, Bowen, GJ, Hayden, B, Pauli, JN, Jordan, C, Anderson, L, Bailey, A, Bataille, CP, Cicero, C, Close, HG, Cook, C, Cook, JA, Desai, AR, Evaristo, J, Filley, TR, France, CAM, Jackson, AL, Kim, SL, Kopf, S, Loisel, J, Manlick, PJ, McFarlin, JM, McMeans, BC, O'Connell, TC, Pilaar Birch, SE, Putman, AL, Semmens, BX, Stantis, C, Stricker, CA, Szejner, P, Trammell, TLE, Uhen, MD, Weintraub-Leff, S, Wooller, MJ, Williams, JW, Yarnes, CT, Vander Zanden, HB and **Newsome, SD.** 2024. 'Design, development, and implementation of IsoBank: A centralized repository for isotopic data'. *PloS one*, 19(9): e0295662. DOI: https://doi.org/10.1371/journal.pone.0295662

Singh, H. 2019. *Practical machine learning and image processing. For facial recognition, object detection, and pattern recognition using python*. 1st ed. Berkeley: Apress. DOI: https://doi.org/10.1007/978-1-4842-4149-3_1

Starbuck, C. 2023. *The fundamentals of people analytics. With applications in R*. 1st ed. Cham: Springer Nature. DOI: https://doi.org/10.1007/978-3-031-28674-2_1

Tanti, M, Berruyer, C, Tafforeau, P, Muscat, A, Farrugia, R, Scerri, K, Valentino, G, Solé, VA and **Briffa, JA.** 2021. 'Automated segmentation of microtomography imaging of Egyptian mummies'. *PloS one*, 16(12): e0260707. DOI: https://doi.org/10.1371/journal.pone.0260707

The pandas development team. 2024. *pandas-dev/pandas: Pandas*. DOI: https://doi.org/10.5281/zenodo.10957263

van den Dries, MH. 1998. *Archaeology and the application of artificial intelligence: case-studies on use-wear analysis of prehistoric flint tools*. Leiden: Faculty of Archaeology, Leiden University.

Viering, T and **Loog, M.** 2023. 'The shape of learning curves: a review'. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(6): 7799–7819. DOI: https://doi.org/10.1109/TPAMI.2022.3220744

Walsh, I, Pollastri, G and **Tosatto, SCE.** 2016. 'Correct machine learning on protein sequences: a peer-reviewing perspective'. *Briefings in Bioinformatics*, 17(5): 831–840. DOI: https://doi.org/10.1093/bib/bbv082

Weiss, GM and **Tian, Y.** 2008. 'Maximizing classifier utility when there are data acquisition and modeling costs'. *Data Mining and Knowledge Discovery*, 17(2): 253–282. DOI: https://doi.org/10.1007/s10618-007-0082-x

Wilkinson, MD, Dumontier, M, Aalbersberg, IJ, Appleton, G, Axton, M, Baak, A, Blomberg, N, Boiten, J-W, da Silva Santos, LB, Bourne, PE, Bouwman, J, Brookes, AJ, Clark, T, Crosas, M, Dillo, I, Dumon, O, Edmunds, S, Evelo, CT, Finkers, R, Gonzalez-Beltran, A, Gray, AJG, Groth, P, Goble, C, Grethe, JS, Heringa, J, 't Hoen, PAC, Hooft, R, Kuhn, T, Kok, R, Kok, J, Lusher, SJ, Martone, ME,

Mons, A, Packer, AL, Persson, B, Rocca-Serra, P, Roos, M, van Schaik, R, Sansone, S-A, Schultes, E, Sengstag, T, Slater, T, Strawn, G, Swertz, MA, Thompson, M, van der Lei, J, van Mulligen, E, Velterop, J, Waagmeester, A, Wittenburg, P, Wolstencroft, K, Zhao, J and **Mons, B.** 2016. 'The FAIR Guiding Principles for scientific data management and stewardship'. *Scientific Data*, 3(1): 160018. DOI: https://doi.org/10.1038/sdata.2016.18

Wolpert, DH. 1996. 'The lack of a priori distinctions between learning algorithms'. *Neural Computation*, 8(7): 1341–1390. DOI: https://doi.org/10.1162/neco.1996.8.7.1341

Xu, P, Ji, X, Li, M and **Lu, W.** 2023. 'Small data machine learning in materials science'. *npj Computational Materials*, 9(1): 42. DOI: https://doi.org/10.1038/s41524-023-01000-z

Yezzi-Woodley, K, Terwilliger, A, Li, J, Chen, E, Tappen, M, Calder, J and **Olver, PJ.** 2022. 'Using machine learning on new feature sets extracted from 3D models of broken animal bones to classify fragments according to break agent'. *arXiv.2205.10430v1*. DOI: https://doi.org/10.48550/arXiv.2205.10430

Yosinski, J, Clune, J, Bengio, Y and **Lipson, H.** 2014. 'How transferable are features in deep neural networks?' *arXiv.1411.1792*. DOI: https://doi.org/10.48550/arXiv.1411.1792

Yravedra, J, Herranz-Rodrigo, D, Mendoza, C, Aragón-Poza, P and **Courtenay, LA.** 2021. 'The use of tooth marks for new research into identifying and understanding the first domestic dogs in Palaeolithic populations'. *Journal of Archaeological Science: Reports*, 40: 103252. DOI: https://doi.org/10.1016/j.jasrep.2021.103252

Zantvoort, K, Nacke, B, Görlich, D, Hornstein, S, Jacobi, C and **Funk, B.** 2024. 'Estimation of minimal data sets sizes for machine learning predictions in digital mental health interventions'. *npj Digital Medicine*, 7(1): 361. DOI: https://doi.org/10.1038/s41746-024-01360-w

Zhang, J, Fang, I, Zhang, J, Wu, H, Kaushik, A, Rodriguez, A, Zhao, H, Zheng, Z, Iovita, R and **Feng, C.** 2024. 'Luwa dataset: Learning lithic use-wear analysis on microscopic images'. In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, pp. 22563–22573. DOI: https://doi.org/10.1109/CVPR52733.2024.02129

Zickel, M, Gröbner, M, Röpke, A and **Kehl, M.** 2024. 'MiGIS: micromorphological soil and sediment thin section analysis using an open-source GIS and machine learning approach'. *E&G Quaternary Science Journal*, 73(1): 69–93. DOI: https://doi.org/10.5194/egqsj-73-69-2024

TO CITE THIS ARTICLE:
Eleftheriadou, A, McPherron, SP and Marreiros, J. 2025. Machine Learning Applications in Use-Wear Analysis: A Critical Review. *Journal of Computer Applications in Archaeology*, 8(1): 188–205. DOI: https://doi.org/10.5334/jcaa.190

Submitted: 29 November 2024 **Accepted:** 18 April 2025 **Published:** 05 June 2025