

# Probability Distribution as an Input to Machine Learning Tasks

Karel Macek<sup>1</sup><sup>a</sup>, Nicholas Čapek<sup>1</sup><sup>b</sup> and Nikola Pajerová<sup>2</sup><sup>c</sup>

<sup>1</sup>AI Center of Excellence, Generali Česká pojišťovna, Na Pankráci 1720, Prague, Czechia

<sup>2</sup>Department of Technical Mathematics, Faculty of Mechanical Engineering, CTU, Resslova 307, Prague, Czechia

**Keywords:** Machine Learning, Classification, Regression, Random Sample, Vectorization, Image Similarity, Hip Bone, 3D Scans.

**Abstract:** Machine Learning has been working with various inputs, including multimedia or graphs. Some practical applications motivate using unordered sets considered to be samples from a probability distribution. These sets might be significant in size and not fixed in length. Standard sequence models do not seem appropriate since the order does not play any role. The present work examines four alternative transformations of these inputs into fixed-length vectors. This paper demonstrates the approach in two case studies. In the first one, pairs of scans as coming from the same document based were classified on the distribution of lengths between the reference points. In the second one, the person's age based on the distribution of D1 characteristics of the 3D scan of their hip bones was predicted.

## 1 INTRODUCTION

Supervised machine learning has been the most fruitful application of artificial intelligence in regression and classification tasks where some input is modeled to predict a quantity or a class, respectively. Both machine learning practice and theory worked not only with vectors of numbers but also (time) series (Ahmed et al., 2010), multimedia (Camastra and Vinciarelli, 2015), graph (Chami et al., 2022). This article discusses the possibility of using a set of measurements as input from a probability distribution. The following examples motivate the topic:

- To model the Gross Domestic Product of a country based on the age structure of its population.
- To identify the type of production machinery's fault based on the distribution of produced product deviations.
- To model a person's age based on the distribution of all points coming from a 3D scan (Kotěrová et al., 2018).
- To predict if two images are the same or not - based on the distribution of distances among the pairs of identified reference points (Čapek, 2022).

According to our knowledge, only a few publications examined the possibility of using a set of measurements from a probability distribution as an input. For example, the authors of (Vinyals et al., 2015) deal with unordered data sets as input for Machine Learning. However, they narrowly focused on sequence-to-sequence mapping and only restricted to Recurrent Neural Networks.

This article provides a general approach to using samples from record-specific probability distribution as input to machine learning tasks. First, we start with the formal problem definition in Section 2. Then, Section 3 proposes the methods to solve it that are demonstrated in two case studies in Section 4 and 5. Finally, Section 6 concludes the article.


## 2 PROBLEM STATEMENT


### 2.1 Supervised Machine Learning Problems


The supervised machine learning can be concisely formulated as optimization of parameters of a model to minimize the prediction error on the training data set (Murphy, 2012, page 179).

More formally, we assume a model

$$p(y|x, \theta)$$

<sup>a</sup> <https://orcid.org/0000-0002-3914-447X>

<sup>b</sup> <https://orcid.org/0000-0002-8513-9540>

<sup>c</sup> <https://orcid.org/0000-0002-7515-3082>

and a training set of input-output pairs  $(x_i, y_i)$  where  $x_i \in \mathbb{R}^n$  and

- for binary classification  $y_i \in \{0, 1\}$  for  $i = 1, 2, \dots, m$
- and for regression  $y_i \in \mathbb{R}$  for  $i = 1, 2, \dots, m$ .

The goal of supervised machine learning is to find such a parameter  $\theta$  that the model predicts based on the available  $x$  the output  $y$  as precisely as possible: in our experiments, accuracy for binary classification and mean absolute percentage error were considered.

## 2.2 Distributional Input

This article addresses a related, yet different problem:  $x_i$  is not a vector but a sample from a probability distribution. Therefore, it has not a fixed length, and the order does not matter, thus  $x_i = \{x_{i,1}, x_{i,2}, \dots, x_{i,n_i}\}$ , where  $n_i$  is the number of observations for record  $i$ .

No additional assumptions on the distribution and the proposed approaches expect only the fact that the input has the above-described properties, i.e., it is an unordered set of numbers.

## 3 METHODOLOGY

The methodology combines suitable informative vectorization, i.e., the transformation of the sets of measurements  $x_i$  for  $i = 1, \dots, m$  into vectors. Subsequently, this vectorized representation can work with standard machine learning models and related evaluation mechanisms.

### 3.1 Vectorization Approaches

Thus, we need to transform the sets of observations  $x_i$  to a fixed-length summary. We propose four alternative ways.

#### 3.1.1 Mean Only Vectorization

In this case, we calculate only the empirical mean of observations for each record:

$$z_i^{\text{avg}} = \left[ \frac{1}{n_i} \sum_{j=1}^{n_i} x_{i,j} \right] \quad (1)$$

The vectorization is one-dimensional. We will use it as a dummy benchmark to indicate that taking the average only implies a loss of discriminative information in the raw data.

#### 3.1.2 Vectorization Based on Empirical Statistics

An extension of the previous approach is to calculate more than one statistic. We consider:

- mean  $\hat{\mu}(x_i)$ ,
- standard deviation  $\hat{\sigma}(x_i)$ ,
- variance  $\hat{\sigma}^2(x_i)$ ,
- minimum  $\min x_i$ ,
- maximum  $\max x_i$ ,
- sum  $\sum x_i$ ,
- kurtosis  $\widehat{\text{Kurt}}(x_i)$ ,
- skewness  $\widehat{\text{skew}}(x_i)$ ,
- and quantiles for levels 10%, 25%, 50%, 75%, and 90%, i.e.  $\hat{q}_{.1}(x_i), \hat{q}_{.25}(x_i), \hat{q}_{.5}(x_i), \hat{q}_{.75}(x_i), \hat{q}_{.9}(x_i)$ .

We represent these values as the following vector:

$$z_i^{\text{stats}} = [\hat{\mu}(x_i), \hat{\sigma}(x_i), \hat{\sigma}^2(x_i), \min x_i, \max x_i, \sum x_i, \widehat{\text{Kurt}}(x_i), \widehat{\text{skew}}(x_i), \hat{q}_{.1}(x_i), \hat{q}_{.25}(x_i), \hat{q}_{.5}(x_i), \hat{q}_{.75}(x_i), \hat{q}_{.9}(x_i)] \quad (2)$$

#### 3.1.3 Binning Vectorization

Another way to represent the distribution in a fixed length form is binning, for example, in (Kotěrová et al., 2018). We define the number of observations in a bin like  $n_k(x_i) = \#\{x_{i,j} | l_k \leq x_{i,j} < u_k\}$  where  $l_k, u_k$  are lower and upper bound of a bin, respectively, and  $k$  denotes the bin's number  $k = 1, \dots, K$ .

$$z_i^{\text{bin}} = [n_k(x_i)]_{k=1}^K \quad (3)$$

We propose the quantile-based approach for defining the bins, i.e., the data set  $x_i = \{x_{i,j}\}$  cut based on quantiles.

#### 3.1.4 Vectorization Based on Likelihood Projections

The following method is motivated by Radial Basis Functions with Gaussians (Scholkopf et al., 1997) where  $x_i$  gets a list of features that correspond to the value of normal probability density function with a specific center and variance.

Assuming that some probability distributions are characteristic of some values of the Machine Learning output  $y$ , we can consider how much the values in  $x_i$  match or do not match a probability distribution. More formally, we consider a set of distributions  $f_d$

for  $d = 1, 2, \dots, D$ . The level to which the data match is quantified as log-likelihood:

$$z_{i,d}^{\text{lik}} = \log L_{f_d}(x_i) = \log \prod_{j=1}^{n_i} f_d(x_{i,j}) \quad (4)$$

This method generalizes the binning approach if we consider the bins as uniform distributions.

The question is how to define the set of characteristic distributions  $f_d$  for  $d = 1, 2, \dots, D$ . We propose the following approach that generates a rich space of them - to consider four normal distributions for each  $i = 1, \dots, n$ :

- $\mathcal{N}(\hat{\mu}(x_i), \hat{\sigma}(x_i))$
- $\mathcal{N}\left(\hat{\mu}(x_i), \frac{\hat{\sigma}(x_i)}{2}\right)$
- $\mathcal{N}\left(\hat{\mu}(x_i) - \frac{\hat{\sigma}(x_i)}{2}, \frac{\hat{\sigma}(x_i)}{2}\right)$
- $\mathcal{N}\left(\hat{\mu}(x_i) + \frac{\hat{\sigma}(x_i)}{2}, \frac{\hat{\sigma}(x_i)}{2}\right)$

Thus, we generate an abundance of  $D = 4 \cdot m$  distributions, which requires a robust regularization approach.

The motivation for this choice of distributions is to capture the each-other matches between  $i$  and  $j$  records for  $i, j \in \{1, \dots, n_i\}$  and whether record  $i$  has values below or above the record  $j$ .

### 3.2 Note on Comparison

When using the introduced vectorizations in machine learning tasks, we considered two approaches:

- **Approach 1:** To combine the vectorization with min-max scaler and a simple model with robust regularization. For example, the logistic regression can be applied with cross-validation to select the right regularization parameter (Golub et al., 1979). Similarly, we can use Lasso for regression. The essential advantage of this approach is the interpretability of coefficients. The robust regularization makes it applicable to all vectorization methods, even if they significantly differ in the number of features.
- **Approach 2:** To use an auto ML library that can handle nonlinearity as well as interaction of features. We consider this for the comparison as the only way due to the different numbers of features. We adopted TPOT (Le et al., 2020).

To obtain a statistically sound comparison of various vectorizations, we adopt CV 5x2 test (Alpaydm, 1999) that is broadly adopted as a tool for comparison of machine learning in general.

## 4 CASE STUDY: IMAGE MATCHING

### 4.1 Case Study Statement

Our selected classification problem is motivated by a document-processing pipeline, which requires operators to check if a pair of scans correspond to the same underlying physical document. In this document-processing pipeline<sup>1</sup>, physical documents are scanned twice:

- once using a mobile phone scanning application
- and a second time on standard office scanners.

We call these mobile scans and standard scans, respectively. Therefore, mobile and standard scans result in near-duplicate but not pixel-perfect, identical scans. Minor differences arise due to lighting, angle, cropping, and differing devices. An example of matching image pairs may be seen in Figure 2a, and non-matching image pairs may be seen in Figure 2b. The task is to determine *whether* a given pair of a mobile scan, and a standard scan are of the same underlying physical document, i.e., a binary target  $y_i$  corresponding to a classification task.

More formally, given two images  $s(d_a)$  and  $s'(d_b)$ , where  $s(d_a)$  is a mobile scan  $s$  of document  $d_a$ , and  $s'(d_b)$  is a standard scan  $s'$  of document  $d_b$ , determine if  $a = b$ :

$$y_i = \begin{cases} 1, & \text{if } a = b \\ 0, & \text{otherwise} \end{cases} \quad (5)$$

Features are extracted using the ORB algorithm (Rublee et al., 2011). The ORB algorithm identifies key points in the image, and each key point has a corresponding feature vector, also known as a descriptor. Keypoints are then matched by pairing key points with the lowest calculated distance between their respective descriptors.

Figure 2 displays key points and their corresponding matches for matching and non-matching image pairs. The top 20 matches are shown. Notice that in Figure 2a, keypoints are matched well but not perfectly, while in Figure 2b understandably, they cannot be matched well. Tendency, but still not sharp clarity, is also evident from Figure 1 where we compare two histograms - one for a case where the scans come from the same document and one where they do not.

Every identified match thus results in a distance based on the quality of the match. The number of identified matches  $n_i$  in each image pair may vary, resulting in a set of observed distances  $x_i$ . Therefore,

<sup>1</sup>More details and business context is described in (Čapek, 2022).

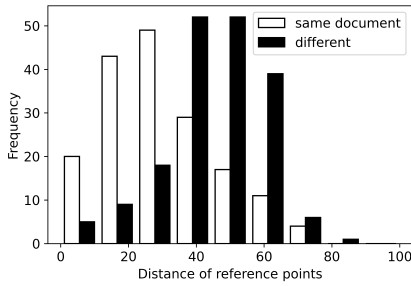


Figure 1: Histogram of  $x_i$  for one case where the scans correspond to the same document (white) and when they differ (black).

we obtain from a pair of images a sample of observations from a distribution and can apply methods in Section 3.

## 4.2 Experimental Results

We worked with 963 pairs of scans from a real-world scanning process in an insurance company. After applying all vectorizations defined in Section 3.1 together with min-max scaling and L1-regularized logistic regression optimized using 5-fold cross-validation, we compared them using the CV 5x2 F-test as introduced in Section 3.2. The results are summarized in Table 4. The columns are the tested approaches. Then, we used also TPOT and the results are in Table 2. The TPOT parameters are in the Appendix.

Table 4 shows that the *mean* method is outperformed by all others that contain more information about the distribution. The same also holds for Table 2. Additionally, the advanced TPOT models prove this case study to be more successful with *stats* than with *lik*. This might be interpreted as the ability of advanced models to interpret the compressed information about the distribution.

Table 1: Results for Image Matching - Logistic Regression. The first part summarizes the results of 10 experiments for CV 5x2, the second part show the results of the pairwise CV 5x2 tests: p-values below significance level  $\alpha = 5\%$  are in bold.

	mean	stats	bins	lik
Mean	63.43%	76.28%	74.35%	74.39%
Std	2.15%	2.01%	1.82%	1.50%
mean	-	<b>0.14%</b>	<b>2.01%</b>	<b>0.48%</b>
stats	<b>0.14%</b>	-	44.03%	12.59%
bins	<b>2.01%</b>	44.03%	-	71.46%
lik	<b>0.48%</b>	12.59%	71.46%	-

Table 2: Results for Image Matching - TPOT. The first part summarizes the results of 10 experiments for CV 5x2, the second part show the results of the pairwise CV 5x2 tests: p-values below significance level  $\alpha = 5\%$  are in bold.

	mean	stats	bins	lik
Mean	62.51%	79.15%	75.97%	77.22%
Std	1.59%	2.15%	1.21%	1.40%
mean	-	<b>0.01%</b>	<b>0.00%</b>	<b>0.01%</b>
stats	<b>0.01%</b>	-	6.25%	<b>1.04%</b>
bins	<b>0.00%</b>	6.25%	-	54.54%
lik	<b>0.01%</b>	<b>1.04%</b>	54.54%	-

## 5 CASE STUDY: HIP BONE AGE PREDICTION

### 5.1 Case Study Statement

For this case study, the data were taken from the optical scanning of hip bones (the collection of 153 scans of female hip bones is taken from (Kotěřová et al., 2022)). Concretely, the part of the hip bone called symphysis was considered for the comparison since it is one part of the hip bone that is used for age determination. Data were in STL format, which means the obtained file is in the form of triangular mesh with face normals (i.e., ordered list of face vertices coordinates followed by the face normal vector). Examples of two female symphyses of ages 21 and 87 are depicted in Figure 3. The detailed structure of the triangular surface for 25-year-old symphysis is depicted in Figure 4. These meshes' parts were aligned into the same position and size (symmetrical according to X- and Y-axis).

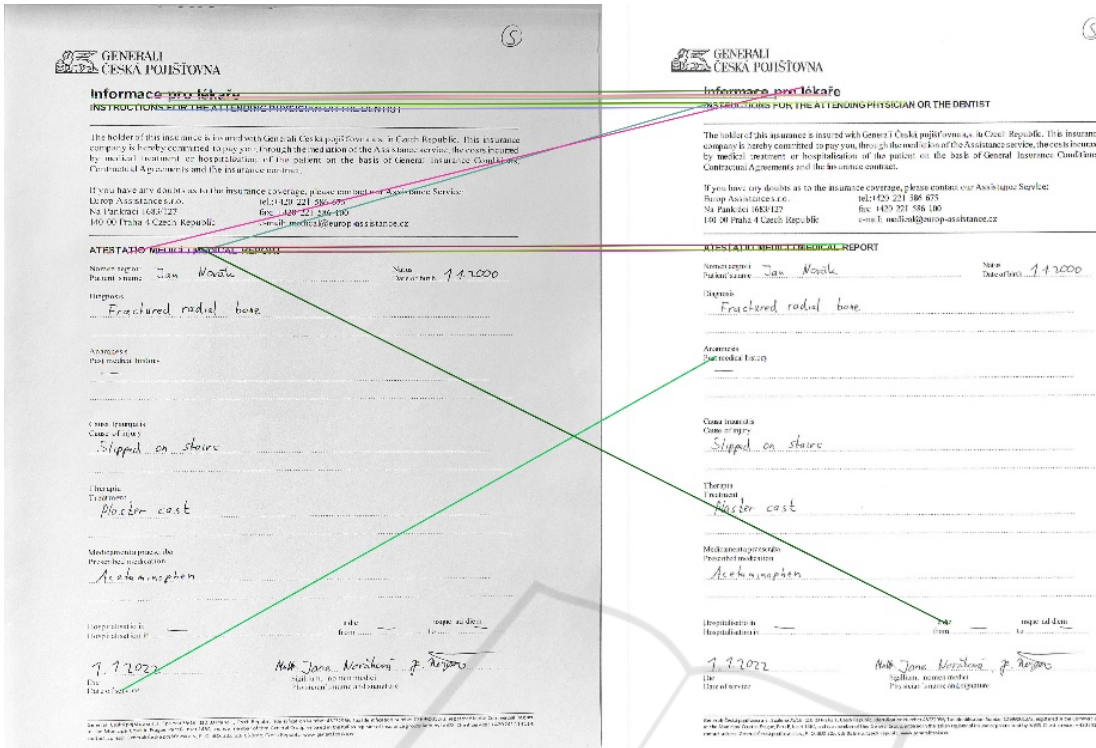
The comparison and age estimation are currently done only visually. However, the difference in surface shapes of different ages can be demonstrated on these two meshes from Figure 3, where the younger bone surface has "furrows" and the older one is more worn.

So the task is to find a sound computer estimation procedure of the age-at-death of the scanned symphysis. To obtain it, the shape function  $D1$  (mentioned in (Osada et al., 2002)) was used, where the function was modified into the discrete version for vertices in this form:

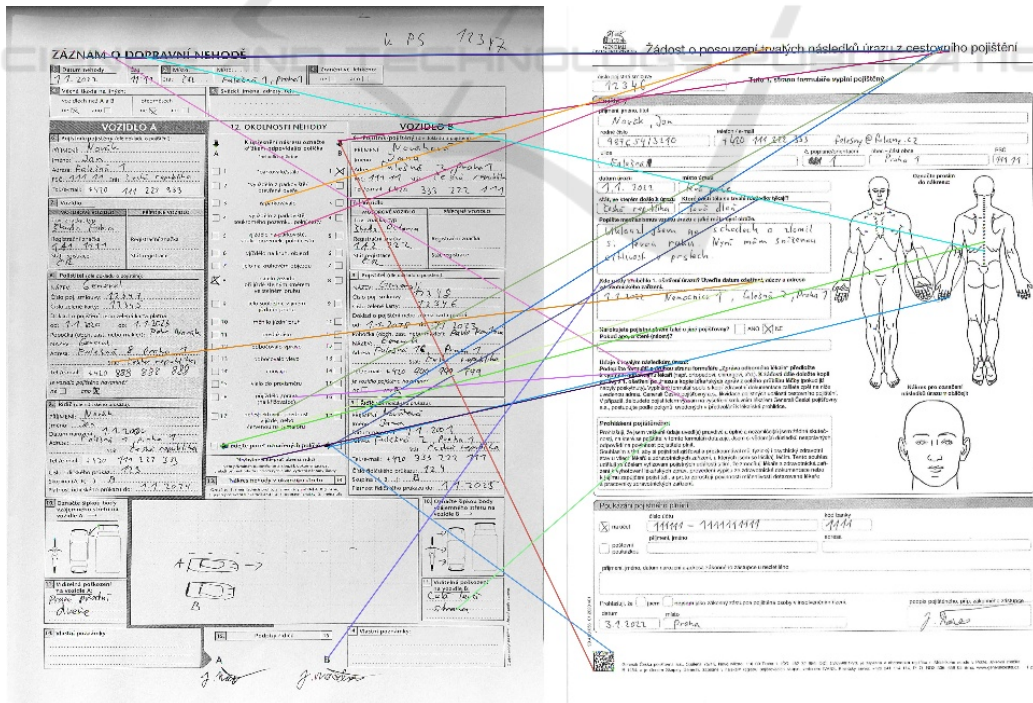
$$x_{i,j} = \text{sign}(v_{i,j,1}) \sqrt{v_{i,j,1}^2 + v_{i,j,2}^2 + v_{i,j,3}^2} \quad (6)$$

where  $v_{i,j} = (v_{i,j,1}, v_{i,j,2}, v_{i,j,3})$  is the  $j$ th vertex in  $i$ th sample with three components that correspond to X, Y, Z axes in 3D. This function measures the oriented distance of mesh vertex from the origin. Note this defi-





(a)



(b)

Figure 2: (a) Image pair with matching physical document with displayed matches (b) Image pair with non-matching physical document with displayed matches.

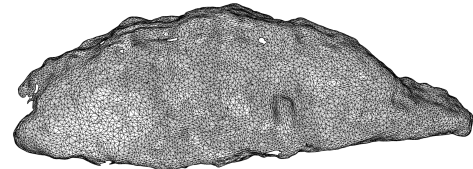
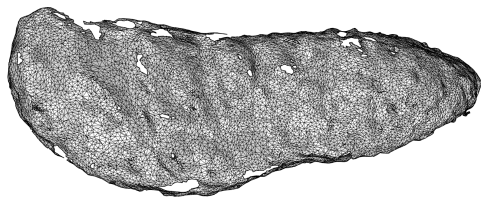


Figure 3: Hip Bone Scan - triangularized vertices.

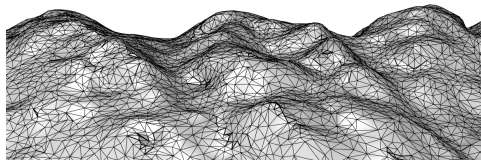


Figure 4: Hip Bone Scan - a detail.

definition of the sign function:

$$\text{sign}(v_{i,j,1}) \begin{cases} 1, & \text{if } v_{i,j,1} \geq 0 \\ -1, & \text{otherwise} \end{cases} \quad (7)$$

Using this function, we can compare the difference of histograms for these two meshes in Figure 5. First, the histogram values are calculated for the given mesh from the resulting data after applying the  $D1$  function. Then the frequencies are normalized to eliminate the effect of a different number of vertices.

## 5.2 Experimental Results

The experimental results are summarized in Table 3 and Table 4 for Lasso regression and TPOT regression, respectively. As we can see, the approaches do not differ significantly, and the more informative vectorizations did not outperform the simple method that

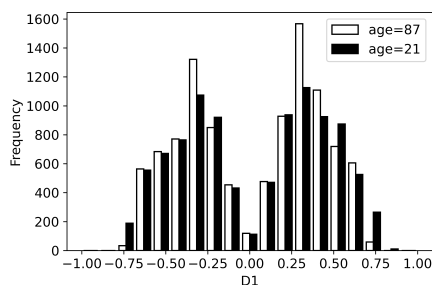


Figure 5: Histogram of  $x_i$  for one woman of 87 years (white) and one of 21 years (black).

uses mean. The configuration of TPOT is in the appendix.

Table 3: Results for Age Prediction with Lasso. The first part summarizes the MAPE (Mean Absolute Percentage Error) results of 10 experiments for CV 5x2, the second part show the results of the pairwise CV 5x2 tests: no p-values are below the significance level  $\alpha = 5\%$ .

	mean	stats	bins	lik
Mean	31.49%	31.64%	31.46%	31.25%
Std	2.24%	2.06%	2.17%	2.26%
mean	-	48.99%	36.05%	10.82%
stats	48.99%	-	47.51%	52.59%
bins	36.05%	47.51%	-	51.26%
lik	10.82%	52.59%	51.26%	-

Table 4: Results for Age Prediction with TPOT. The first part summarizes the MAPE (Mean Absolute Percentage Error) results of 10 experiments for CV 5x2, the second part show the results of the pairwise CV 5x2 tests: no p-values are below the significance level  $\alpha = 5\%$ .

	mean	stats	bins	lik
Mean	28.00%	30.43%	28.71%	32.43%
Std	1.68%	2.70%	1.77%	2.26%
mean	-	55.17%	10.32%	79.72%
stats	55.17%	-	30.42%	41.19%
bins	10.32%	30.42%	-	10.46%
lik	79.72%	41.19%	10.46%	-

## 6 CONCLUSION

We have examined the supervised machine learning problem with samples from record-specific probability density as an input. We proposed four approaches and compared them to each other in two real-world case studies. Methods that work with a richer representation (empirical statistics, bins, likelihood) outperformed the naive method based on empirical mean statistically in one of the case studies. Moreover, these methods do not differ when used with penalized linear methods (Lasso, Logistic Regression). When using more advanced modeling with the TPOT library, the method based on likelihoods has been outperformed by the method with empirical statistics.

The positive result motivates further research in the area. The vectorization methods can be tested to see significant features, and the possibilities to interpret the models can be further investigated. Another dimension can be examining methods related to Long Short-Term Memory (LSTM) Networks or Probabilistic Graphical Models.

## REFERENCES

- Ahmed, N. K., Atiya, A. F., Gayar, N. E., and El-Shishiny, H. (2010). An empirical comparison of machine learning models for time series forecasting. *Econometric reviews*, 29(5-6):594–621.
- Alpaydm, E. (1999). Combined  $5 \times 2$  cv f test for comparing supervised classification learning algorithms. *Neural computation*, 11(8):1885–1892.
- Camastra, F. and Vinciarelli, A. (2015). *Machine learning for audio, image and video analysis: theory and applications*. Springer.
- Čapek, N. (2022). *Digital Document Analysis Using Machine Learning Methods*. Master’s thesis, Masaryk University, Faculty of Informatics, Brno, Czechia.
- Chami, I., Abu-El-Haija, S., Perozzi, B., Ré, C., and Murphy, K. (2022). Machine learning on graphs: A model and comprehensive taxonomy. *Journal of Machine Learning Research*, 23(89):1–64.
- Golub, G. H., Heath, M., and Wahba, G. (1979). Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics*, 21(2):215–223.
- Kotěřová, A., Navega, D., Štepanovský, M., Buk, Z., Brůžek, J., and Cunha, E. (2018). Age estimation of adult human remains from hip bones using advanced methods. *Forensic Science International*, 287:163–175.
- Kotěřová, A., Štepanovský, M., Buk, Z., Brůžek, J., Techataweewan, N., and Velemínská, J. (2022). The computational age-at-death estimation from 3d surface models of the adult pubic symphysis using data mining methods. *Nature*, Scientific Reports 12.
- Le, T. T., Fu, W., and Moore, J. H. (2020). Scaling tree-based automated machine learning to biomedical big data with a feature set selector. *Bioinformatics*, 36(1):250–256.
- Murphy, K. P. (2012). *Machine Learning: A Probabilistic Perspective*. The MIT Press.
- Osada, R., Funkhouser, T., Chazelle, B., and Dobkin, D. (2002). Shape distributions. *ACM Transactions on Graphics (TOG)*, 21(4):807–832.
- Ruble, E., Rabaud, V., Konolige, K., and Bradski, G. (2011). Orb: An efficient alternative to sift or surf. In *2011 International conference on computer vision*, pages 2564–2571. Ieee.
- Scholkopf, B., Sung, K.-K., Burges, C. J., Girosi, F., Niyogi, P., Poggio, T., and Vapnik, V. (1997). Comparing support vector machines with gaussian kernels to radial basis function classifiers. *IEEE transactions on Signal Processing*, 45(11):2758–2765.
- Vinyals, O., Bengio, S., and Kudlur, M. (2015). Order matters: Sequence to sequence for sets. *arXiv preprint arXiv:1511.06391*.

## APPENDIX

### TPOT Configuration

TPOT an auto-ML Python library<sup>2</sup> that was used in the experiments for more advanced modeling. For the classification, the following configuration was used:

```
TPOTClassifier(generations=5,
               population_size=20,
               cv=5,
               random_state=42,
               verbosity=2)
```

For the regression, the following configuration was used:

```
TPOTRegressor(
    generations=5,
    population_size=20,
    cv=5,
    random_state=42,
    verbosity=2,
    scoring=make_scorer(
        mean_absolute_percentage_error,
        greater_is_better=False))
```

<sup>2</sup><http://epistasislab.github.io/tpot/>