Improving Color Mixture Predictions in Ceramics using Data-centric Deep Learning

Tomas Souper
tomas.pereira@aicos.fraunhofer.pt
Fraunhofer Portugal AICOS
Porto, Portugal

Ana C. Morgado
ana.morgado@aicos.fraunhofer.pt
Fraunhofer Portugal AICOS
Porto, Portugal

Ana Marques
anamarques@matceramica.com
Matcerâmica - Fabrico de Louça S.A.
Batalha, Portugal

Inês Silva
laboratorioa@matceramica.com
Matcerâmica - Fabrico de Louça S.A.
Batalha, Portugal

Luís Rosado
luis.rosado@aicos.fraunhofer.pt
Fraunhofer Portugal AICOS
Porto, Portugal

ABSTRACT

Ceramics is a millenary industrial sector with relevant financial impact for several countries. Efficiency in color mixing is crucial in the ceramic industry, both in terms of staff time and consumables costs. Traditional color mixing methods usually consist of manual processes based on in-depth domain knowledge of basic color theory or color models like Kubelka-Munk. Thus, the efficiency of these procedures is highly dependent on the technician's expertise, being challenging for novices to acquire these skills and be proficient. This work explores the usage of Deep Learning to generate color mixture predictions in ceramic glazes. The proposed solution is based on spectral data of ceramic components (pigments and glazes) and, based on their respective quantities, simulates the color mixing result in a wholly digital way. Given the lack of freely available datasets, we started by exploring our approach in the NTU Watercolor Pigments Spectral Measurement dataset. We then translated the collected knowledge to the Matcerâmica Ceramics Spectral Measurement dataset, which was specifically created in the ambit of this work. By using data-centric optimization techniques to improve our model interactively, the best performance was achieved by fully connected neural network model, with mean ΔE^*_{ab} of 1.57 and 2.35 for the NTU and Matceramica datasets, respectively. These results demonstrate the potential of the proposed approach to be integrated into an AI-powered software solution that improves color mixing procedures in the ceramic industry.

CCS CONCEPTS

• **Computing methodologies** → **Neural networks**; *Artificial intelligence*; • **Applied computing** → *Computer-aided manufacturing*.

KEYWORDS

Deep Learning, Neural Networks, Color Mixture Prediction, Ceramics

ICMLT 2023, March 10–12, 2023, Stockholm, Sweden

ACM ISBN 978-1-4503-9832-9/23/03.
https://doi.org/10.1145/3589883.3589917

ACM Reference Format:
Tomas Souper, Ana C. Morgado, Ana Marques, Inês Silva, and Luís Rosado. 2023. Improving Color Mixture Predictions in Ceramics using Data-centric Deep Learning. In *2023 8th International Conference on Machine Learning Technologies (ICMLT 2023), March 10–12, 2023, Stockholm, Sweden.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3589883.3589917

1 INTRODUCTION

Ceramics is a relevant industrial sector in several countries, translated into a high global volume of business, especially in exports. Despite having a millenary tradition, ceramic producers have continuously innovated and maintained their competitiveness worldwide. In particular, digital transformation in the ceramic industry has the potential to improve the efficiency and productiveness of several processes, including color mixing procedures. Whether in development or industrial production phases, reliability and cost-efficiency in the development of colors are essential to ensure competitive advantages. Traditional color mixing methods involve manual processes based on in-depth domain knowledge like basic color theory. To achieve a target color, technicians must find the ideal set of components to mix and their corresponding quantities. Thus, the efficiency of these procedures is highly dependent on the technician's expertise, being challenging for novice technicians to acquire these skills quickly. To tackle the bottlenecks of traditional methods, digital and simulation models are finding increased use in other industry segments like the paint, plastic, or dyed textile industries. These systems are based on color measurements (e.g. via spectrometers, spectrophotometers, or colorimeters) and aim to reduce current inefficiencies, namely by replacing manual trial-and-error methods with efficient digital test systems.

With this in mind, computer-aided systems to support color-mixing procedures have started to appear. Most of the proposed approaches are based on the color model created by Kubelka-Munk in 1931 [11]. This model uses non-linear equations that relate the absorption and scattering of each component of a color mixture to obtain the mixture reflectance. Instead of producing every possible recipe, this model may be used to simulate the outcome of a wide range of color mixtures and consequently narrow the number of recipes that are then produced and validated in real-world conditions. In recent years, the Kubelka-Munk model has started to be replaced by more complex models, with Neural Networks (NN) starting to achieve better results than the traditional methods. This

work explores the usage of Deep Learning to generate color mixture predictions in ceramic glazes. The proposed solution is based on spectral data of ceramic components (pigments and glazes) and, based on their respective quantities, simulates the color mixing result in a wholly digital way.

The paper is structured as follows: Section 1 provides the motivation and objective of the work; Section 2 summarizes the relevant related work; Section 3 details the used datasets; Section 4 describes the proposed methodology; Section 5 details the results and the respective discussion; and Section 6 draws the conclusions and future work.

2 RELATED WORK

This section details the relevant work in the literature regarding color mixture prediction, namely in terms of Traditional Methods and Machine Learning Methods.

2.1 Traditional Methods

Color measurement is an important and complex process that depends on human eye perception and is related to the visible region of the spectre. A color space is an organization of colors that aims to support reproducible representations of color, which can be analog or digital. The CIELAB color space, also known as L*a*b*, tries to express the color as three values: L* (perceptual lightness); a* (red/green coordinate); and b* (yellow/blue coordinate). This color space tries to be a perceptually uniform space where a numerical change can be mapped to a perceived change in color. Color difference is a metric that tries to quantify the separation between two colors. In the L*a*b* color space the metric is called ΔE^*_{ab} and due to some lack of perceptual uniformity, specially in the blue hues, is given by 1:

$$\Delta E^*_{ab} = \sqrt{\left(\frac{\Delta L'}{k_L S_L}\right)^2 + \left(\frac{\Delta C'}{k_C S_C}\right)^2 + \left(\frac{\Delta H'}{k_H S_H}\right)^2 + R_T \frac{\Delta C'}{k_C S_C} \frac{\Delta H'}{k_H S_H}} \tag{1}$$

where ΔL, ΔC and ΔH are the differences in lightness, chroma and hue; k_L, k_C, and k_H are domain-dependent parametric factors; S_L, S_C, and S_H correspond to weighting functions for the lightness, chroma, and hue components, respectively; and R_T consists of a rotation function to compensate the lack of perceptual uniformity in the blue region [12]. Due to this properties this metric tries to quantify the difference of two colors perceived by the human eye. If has a $\Delta E^*_{ab} <= 2$ the human eye cannot perceive difference between colors. In some colors this threshold can be 5. If the difference is higher than 5 the human eye will clearly see the difference.

Kubelka-Munk (K-M) theory tries to model the incident behavior of a mixture of components using its absorption and scattering. K-M equation describes the reflectance with the coefficient of absorption and the scattering coefficient on each wavelength of the visible region of the spectre.

This relationship can be applied to different opaque products including opaque ceramics [14]. Based on the Kubelka Munk equation, it was later demonstrated the additive properties of individual contributions of absorption and scattering in a mixture. This is called Duncan Equation and describes the absorbed light due to pigment mixture by using the concentrations of each component and their scattering and absorption coefficient. Different industries use these equations to predict the reflectances of a mixture with different pigments. In the ceramics industry, some different studies have been made. In study [3], both Kubelka-Munk and Duncan equations zwere used. The results showed good predictions in some regions of the visible spectrum, concluding that the model could be used for quality control of products manufactured in the ceramic industry. The work published in [14] uses the transformed the K-M equation, by splitting in three components one for the glaze, another for the opacifier and a final one for the pigment. The authors concluded that color reproduction was possible with a reduced number of experiments and the new model made it possible to correlate the colour with the added components.

2.2 Machine Learning Methods

Even though neural networks grew in popularity in the last decade they have been used to tackle this problem for some time [1]. These networks try to mimic the behaviour of the human brain. There are input values that are combined with each other in order to obtain output values. It works like a network where input values are multiplied by weights and summed together, then it passes through an activation function that will introduce non-linearity. In the beginning, the weights are random but through a process of training the weights are corrected using backpropagation by trying to minimise the prediction error. Neural networks started to achieve better results than traditional methods in the task of recipe prediction [2]. The study [6] developed a neural network to predict the reflectances of a mixture of fibres. The input values are a weighted sum of the reflectances values of each component with its concentration. The neural network was capable of providing good results which reduced the total time for optimizing a colour recipe. Similar to the previous study, in [10] instead of trying to predict the reflectance values, the neural networks tried to predict color in space XYZ. The results show that neural networks could achieve better results than the conventional Kubelka-Munk model. The work of [8] uses the concentrations from the 8 primaries colors as input. However, as output there are two different approaches, the first one, similar to other studies, uses a neural to predict all (35) reflectance values. The second approach was to use a neural network to predict each single value of reflectance. In both cases, it achieved better results than the traditional Kubelka-Munk method. Even though in some studies neural networks start to achieve better performance than traditional methods, study [7] uses a combination between the neural networks and the Kubelka-Munk model. Instead of using the neural networks to predict the color values such reflectance, XYZ or L*a*b* this work uses the neural networks to obtain a closer approximation of the ratio of absorption and scattering (K-S). This ratio is then used to obtain the reflectance values of the mixture. In [5], the authors proposed a neural network to predict the reflectance values of a mixture of two pigments. For this, they created a dataset (the NTU Watercolor Pigments Spectral Measurement dataset) with the values of transmittance and reflectance of 14 pigments that formed mixtures of two pigments, which were sampled in 12 quantities. The spectral data was obtained using a spectrophotometer, which retrieved spectral values from 380nm-780nm with an interval of 10nm. For each pigment, the relative

transmittance was obtained, as well as the absolute reflectance. Two types of data were used: the first was the addition of one pigment to itself increasing the amount of paint used; the second was to mix two pigments with different mixing ratios: (1:1, 1:2, 2:1). The inputs of the neural network then comprised the transmittance, reflectance, quantities of each pigment and the reflectance of the paper where the mixture was applied. With respect to the outputs, they consisted of the reflectance of the resulting color mixture. Delta E was used to measure the color difference between the prediction and the ground truth. Also in this case, the results achieved with the proposed model were better than the results achieved with the traditional Kubelka-Munk model.

In study [15], two spectral datasets were also developed from different paint pigments. In the first, 8 different pigments were combined in order to obtain 286 samples from oil paints. In the second set 397 samples were obtained from 9 different watercolor paints. The authors chose a limited set of 8/9 paints because it is the amount of colors that the most prominent artists use. The black color was not used, so the dark colors were achieved by mixing complementary colors. Water was used to dilute and lighten up the samples, and was added in relatively very large quantity to achieve the light values. To obtain the spectral data, a spectrophotometer was used to sample spectral values from 400nm to 700nm with a interval of 10nm, in order to explore the "true" spectral dimensionality of the paint datasets. Compression and decompression of these datasets was made using the well-known Principal Component Analysis (PCA) method. After this process, the authors concluded that 7 dimensions was enough to reconstruct the color with an error (CIEDE2000) less than 0.5.

Regarding the ceramic industry, to the best of our knowledge, the exploration of neural networks to solve color-mixture tasks is still very scarce. One of the few exceptions is the work of [16], which uses neural networks to build a ceramic color perceptual semantic mapping model, using color quantified values as the input layer and L*a*b* as the output layer. Even though the work is also in the area of color science in ceramics, the end goal is not to make color mixtures, but to solve to solve the problem of mapping between ceramic color appearance characteristics and perceptual semantics.

In summary, it is clear that machine learning approaches for color mixture prediction brought significant improvements. However, research on how to use NN-based approaches to improve color mixture predictions in ceramic glazes is still missing. Additionally, given that spectral data collection for ceramic pieces can be extremely time-consuming and costly, the usage of data-centric ML and synthetic data generation strategies is a topic that remains unexplored in this field.

3 DATASETS

3.1 NTU Watercolor Pigments Spectral Measurement dataset

In the first part of this work, the NTU Watercolor Pigments Spectral Measurement, hereinafter referred to as WSPM dataset, was considered [5]. As previously introduced in Section 2.2, this dataset uses the spectral data of the pigments and substrate where they are being applied (white paper) in order to simulate the color resulting from mixing them. Therefore, the inputs comprise 41 values corresponding to the transmittances and reflectances of each pigment (2*(41+41)), 41 values that represent the paper's reflectances, and the quantities of each pigment. As labels, the authors considered the reflectances of the resulting mixture. In our work, we decided to make some adjustments to this dataset. Instead of directly using the quantities of each pigment as provided by the original dataset, these were transformed into percentage values. Moreover, since in ceramics we cannot obtain transmittance values, we did not include them in our study. After this process, we ended with a total of 125 features - 41 * 2 representing the reflectances of each pigment, 41 values of the reflectance of the paper, and the percentages of each pigment.

With respect to the train/test splitting, we used the original division proposed in [5]. This translates into a total of 1956 color mixtures for training and 488 for testing, which are presented in Figure 1. It is, however, noteworthy that during our experiments we applied some data augmentation techniques, as will be further discussed in the paper.

Figure 1: Color mixtures that form the WPSM testing dataset.

3.2 Matcerâmica Ceramics Spectral Measurement dataset

Since this work aims to develop a system able to predict color mixture predictions in ceramic glazes, we created a dataset called Matcerâmica Ceramics Spectral Measurement (CSM) to simulate color formulations for the ceramic industry. Each color formulation comprises the used components and respective quantities, whereas the targets consist of the L*a*b* value of the color resulting from mixing those components. For each component there is a reading for spectral data and L*a*b* values.

The two main components of each mixture are glazes and pigments. The glaze is a liquid where the pigments are going to be mixed and which is then applied to the ceramics piece. In this study, we used two types of glazes: transparent (stronger colors) and opaque (which result in less vividly colors) Since these components contain mineral elements in their composition, their properties might slightly differ among diverse batches (lots). This means that for a single component we may have multiple readings, as different properties affect the color in different ways, e.g. if we have received two batches of cobalt blue, we will have readings for each one of them. The process of acquiring the readings (spectral data and L*a*b* values) for glazes and pigments is different. For the glazes,

the glaze is directly applied to a plate before going into the oven. In the case of the pigments, the process is a bit more tricky because they have to be mixed with a glaze before being applied to the plates. As we used two types of glazes, for each pigment, we obtained the readings in a transparent glaze and in an opaque one. In both cases (glazes and pigments), the readings were only obtained after the pieces had gone into the oven.

Having this in mind, we explored a total of 205 recipes (color formulations) and 46 unique components. Nevertheless, due the peculiarity of different lots for the same component, this number increases to a total of 308 possible components (as we have a spectral reading for each single lot). Moreover, while in the WSPM dataset every recipe was composed of two components, in the CSM dataset, the amount of components used in each recipe may be different. Thus, to overcome this issue, we performed a "zero padding" until the maximum number of components (11), ensuring that the number of features for each color recipe was the same. Regarding the train/test splitting we used a proportion of 90:10, meaning that the test set is composed of 21 color mixtures.

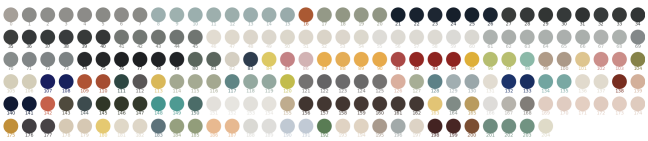


Figure 2: Color mixtures that form the Matceramica dataset.

4 METHODOLOGY

The main objective of this work is to explore the usage of Deep Learning to generate color mixture predictions in ceramic glazes. A color mixture recipe is a combination of components and respective quantities. Thus, the proposed solution is based on spectral data of ceramic components (pigments and bases) and aims to simulate the color mixing result in a wholly digital way. Given the lack of freely available datasets and the effort required to create a new dataset of spectral measurements for ceramic glazes, we started by exploring our approach in the WSPM dataset. This dataset allowed us to try and explore a wider range of machine learning strategies in an initial phase and assess which were the most promising ones. We then translated the collected knowledge to the Matcerâmica Ceramics Spectral Measurement dataset, specifically created in this work's ambit. Even though the datasets are from different industries (paints and ceramics), the core problem is the same. This problem can be defined as a multi-target regression since it tries to predict multiple values in a continuous space.

To tackle this problem, we opted to explore and compare two approaches: i) A Traditional method based on the Kubelk-Munk model; and ii) a Machine Learning method based on a simple Deep Learning model (fully connected). To achieve the best-performing deep learning model, we considered two groups of optimization strategies: model-centric and data-centric. Each group comprises different types of training optimizations iteratively executed and compared. It should be noted that the K-M process does not need training since it is merely a combination of the physical values of each component of the color mixture recipe.

4.1 Model-centric

In the model-centric phase, we iteratively improved the DL model while the datasets were fixed. The main objective was to achieve the best-performing model by fine-tuning the DL models' hyperparameters. In particular, we considered the following parameters and respective range of values:

- **Number of epochs**: 100, 500, 1000, 5000.
- **Batch size**: 32, 64.
- **Model depth**: 4 layers, 6 layers.

The loss function is also a hyperparameter with a relevant impact on the model-centric optimization process. While trying to reproduce the results from [9], we realized that instead of using the RMSE as a loss function (typical regression loss function), the usage of the ΔE^*_{ab} could also be explored as the loss function. Figure 3 illustrates the considered 4-layers fully-connected neural network with two hidden layers. The input layer receives each component's reflectances and percentages, while the output layer directly maps to the target L*a*b* color values.

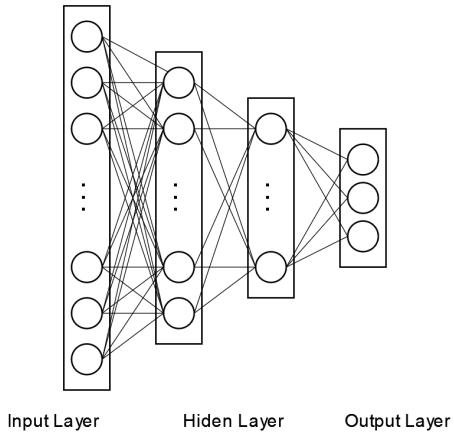


Figure 3: The considered 4-layers fully-connected neural network with two hidden layers.

4.2 Data-centric

In the data-centric phase, we iteratively improved the data quality while the DL model is fixed. By using iterative data-oriented optimization steps that enhance the quality of the train data (rather than collecting more data), we aimed to increase the performance of the most promising DL model found in the model-centric phase. In particular, we explored the following data-centric optimizations steps:

- **From quantities to ratios and percentages** (WSPM, CSM): The quantities of each component can be replaced by ratios or percentages. While in WSPM dataset we only transformed in ratios between 0-1, in CSM dataset we also explored percentages between 0-100%.
- **Data augmentation using Synthetic Data Vault** (WSPM): When analyzing the color space we might find some regions that are less covered. To overcome this limitation we can use synthetic data generation tools to "fill the gaps" in the

color space. We used the Synthetic Data Vault [13] library, which includes state-of-the-art deep learning techniques to generate synthetic data like conditional-tabular generative adversarial networks (CT-GANS) and tabular variational auto encoders (TVAE).

- **Data augmentation using domain knowledge**:
  - **New recipes with only one component with 100% percentage** (WSPM, CSM): We know that in a color recipe when we have a component with 100%, the result of this mixture has to be exactly the color of that component. Thus, we added new instances from existing recipes where one of the components has 100% while the others have zero, and the target is the component's color with 100% (we term this data augmentation technique of *self_augmentation*).
  - **New recipes from each component** (WSPM, CSM): We added new instances where the reading of a component was transformed into a recipe, i.e. the recipe is only composed of a single component with 100% and the input and output reading are similar (we call this data augmentation technique of *component_augmentation*).
  - **New recipes with shuffled components** (CSM): The order of the mixing does not affect the final color. Thus we created new instances by shuffling the order of the components of existing recipes.
- **Data manipulation using domain knowledge:**
  - **Order components by their proportions** (CSM): The bigger the amount of a component in a mixture, the higher the impact of that component on the final color. Thus we added a pre-processing step that orders the components of each recipe by their proportions (descending order from left to right).
  - **Weighted sum of components' opaque and transparent glaze readings** (CSM): Instead of using only opaque glaze readings for each component, for each nm value of the spectral data use the weighted sum between opaque and transparent readings (according to the percentage used for that specific component).
  - **Unfold components' opaque and transparent glaze readings** (CSM): Unfold each component reading into two components: with 100% opaque glaze and 100% transparent glaze. The relative amount of the respective glaze in the recipe defines the quantity percentage of each unfolded component.

4.3 Evaluation

To evaluate the studied models, we used the average value of the loss function. It should be noted that we explored the usage of two loss functions, namely RMSE and ΔE^*_{ab}. After some experiments further detailed in Section 5, it became clear that the best criteria to select the best-performing model was the mean ΔE^*_{ab}. This finding might be supported by the fact that the L*a*b* colorspace tries to mimic the behavior of the human eye.

Additionally, we aimed to assess the ratio of color mixture recipes in the test set with predictions that could be accepted as correct predictions. For that purpose, we used thresholds for the ΔE^*_{ab} values obtained for each test instance, for which a value below the threshold indicates a correct prediction. For the WPSM dataset, we used $\Delta E^*_{ab} < 5$, which was the threshold used in the original paper [5]. For the CSM dataset, we used both used $\Delta E^*_{ab} < 5$ and used $\Delta E^*_{ab} < 2$, since we wanted to consider a more restrictive evaluation environment, given the feedback received from ceramic technicians that indicated that a threshold of 5 might not be sufficient to accept a prediction as correct.

As a final note, to facilitate the comparison of the different experiments and to find the most suitable combination of model-centric and data-centric optimizations, we used MLflow [4], an experiment tracking tool that allows logging the used combinations of hyperparameters, metrics, and even the model itself during the model training process.

5 RESULTS AND DISCUSSION

5.1 NTU Watercolor Pigments Spectral Measurement dataset

As stated in Section 2.1, the K-M equation is an analytic method used to predict the reflectance of a mixture of pigments. We decided to use this method as a baseline for our study. Since this method merely uses the reflectances and concentrations of the mixed components, we only considered 84 features (41*2 values of reflectances, and 2 values corresponding to the ratio of each component's). Figure 4 provides an illustrative example of using the K-M equation to predict the result of a color mixture. The figure depicts four different color circles: the first two circles represent the pigments that are being mixed; the third circle represents the color predicted by the K-M equation; and the last circle provides the target color. To create this visualization, the reflectances were converted to the RGB color space.

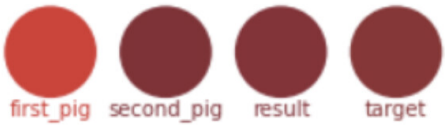


Figure 4: Illustrative example of Kubelka-Munk color mixture prediction in the WSPM dataset: The first two circles are the components of the recipe; the third is the predicted color; and the last circle is the target color (ground truth).

Since the WSPM dataset already comprises train and test sets provided by the authors, we used the latter to retrieve the evaluation of this method, which will be used as our baseline. Two metrics were explored to evaluate the performance: RMSE and the percentage of test instances with $\Delta E^*_{ab} < 5$. With the first iteration using this model, we obtained an average RMSE of 0.003, and the model correctly predicted 52.87% of the test samples ($\Delta E^*_{ab} < 5$), and achieved an average ΔE^*_{ab} of 5.89 in the test set. Since in this dataset we also have access to the reflectance values of the paper where the paint is applied, we decided to use this information inside the model. To do so, we added the paper as a component of the mixture. Figure 5 demonstrates the performance of the K-M equation when using different quantities of paper as a pigment. As it is possible

to verify, when the amount of paper is increased, the performance decreases (RMSE increases), and the amount of predictions with $\Delta E^*_{ab} < 5$ decreases. In the end, the best results achieved did not contemplate the effect of the paper.

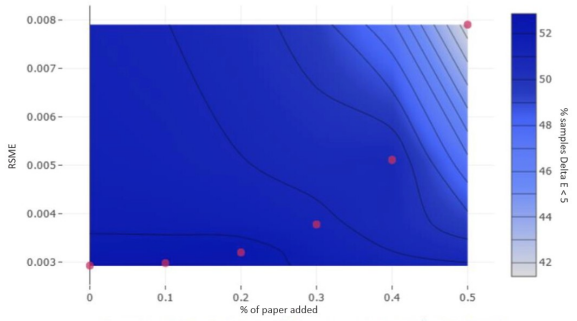


Figure 5: Kubelka-Munk evaluation with paper as pigment in the WSPM dataset. The different colors on the vertical bar represent the amount of samples in the test set that have a ΔE^*_{ab} smaller than 5.

Moving from Traditional to Machine Learning methods, the study that presented the WSPM dataset[5], also proposed DL methods for color mixture prediction. Thus, we started by replicating the results stated in that work: we created a linear model with 4 layers, that has as input the original 207 features and as output 41 values which represent the reflectances of the mixture. For the training of this model, we used a learning rate of 0.1 and a batch size of 32. As presented in Table 1, as we increase the number of epochs the performance improves. The best experiment resulted in a total of 84.43% predicted test samples with a ΔE^*_{ab} less than 5. This result outperforms the one achieved by the simpler K-M equations.

Table 1: Results for the WPSM dataset using the model proposed in [5].

Epochs	Loss Function	Avg Loss	% samples ΔE^*_{ab} <5
100	MSELoss()	0.001	31.56%
500	MSELoss()	8.431e-4	70.49%
1000	MSELoss()	4.331e-4	80.12%
5000	MSELoss()	3.487e-4	84.43%

The authors of the WPSM dataset also released another work [9] where instead of using a linear model, a CNN model based on a ResNet architecture was employed. Beyond this new architecture, the authors also used a different loss function based on eye perception of the color. After implementing these new models, unfortunately, we could not replicate the results presented in [9], only achieving a maximum of 65.57% of good predictions. After analyzing the results, we decided to optimize the total correct predictions with $\Delta E^*_{ab} < 5$. Therefore, instead of using the MSELoss as loss function, we started using the ΔE^*_{ab} function as the loss function. This change allowed the model to optimize the color simulation for the perception of the human eye. To use this loss function, we made some transformations to the dataset, namely the conversion of the target values from reflectances to L*a*b* values. The results in Table 2 demonstrate that transforming both the target values and the loss function leads to better results in terms of the amount of predictions with a $\Delta E^*_{ab} < 5$.

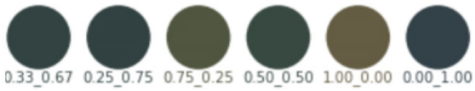
Table 2: Results for the WPSM dataset using different loss functions and target values for 5000 epochs.

Loss Function	Avg Loss	% samples ΔE^*_{ab} <5	Target Values
MSELoss()	3.847e-4	84.43%	Reflectances
MSELoss()	2.353	89.55%	L*a*b*
CIE_2000()	1.905	95.9%	L*a*b*

In order to improve the performance of the best-performing model found during this model-centric stage, we then moved to data-centric optimizations. In particular, we assessed the impact of the following data-driven transformations: converted the quantities of each pigment to ratios; used only pigment reflectances with quantities' ratios; and used pigment and paper reflectances with quantities' ratios. In Table 3, we can see the evolution of the performance with these transformations.

Some illustrative examples are depicted in Figure 6a, where the first two circles represent the pigments we want to mix together, and the last circle represents the result of the mixture between the two pigments.

(a) Recipe from the WSPM dataset



(b) Model predictions for different pigment quantities

Figure 6: On the left, we have a simple recipe from the WSPM dataset. The right side image are predictions made with the best model to date. The values below each circle represent the relative quantity of each pigment. From the domain knowledge, we know that when we have 100% of one pigment, the result of the mixture should that pigment.

The circles contained in Figure 6b represent the prediction of the mixture presented in Figure 6a using different percentages of each pigment. The first circle represents a mixture of 33% orange and 67% blue, and the last one represents 0% orange and 100% blue. From this definition, in the last two circles we should see the color of which has the pigment has 100% of concentration. From our intuition a mixture in which one of the pigment is 100% should output the pigment with 100%. To improve these results, we augmented the dataset with this domain knowledge. We added entries to the dataset where in the input we would have a mixture where one of the pigments had 100% and the target of this entry

Table 3: Results of different data-centric optimizations while keeping epochs as 5000.

Loss Function	Avg Loss	% samples ΔE^*_{ab} <5	Changes
CIE_2000()	1.762	94.26%	Replaced quantities with ratios
CIE_2000()	1.699	96.11%	Only pigment reflectances + ratios
CIE_2000()	1.566	97.13%	Pigment reflectances + paper reflectances + ratios

would be the L*a*b* of this same pigment. After performing this data augmentation, we retrained the model achieving a percentage of correct predictions with $\Delta E^*_{ab} < 5$ of around 95%. In Figure 7, we can see the results for two recipes on edge cases where each pigment has a percentage of 100%.

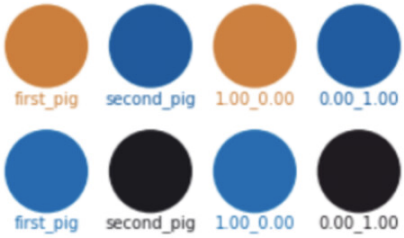


Figure 7: Examples of edge cases on the WSPM dataset after data augmentation (new recipes with only one component with 100% percentage). We can observe that the new model can correctly predict the recipes that only use one pigment.

While the results are incrementally improving with the data-centric optimization steps, when analyzing Figure 8, we can see some regions in the L*a*b* color space without instances representation. Therefore, we decided to try to generate synthetic data in order to fill in the missing spaces. We started to look for oversampling techniques, however, since our problem can be framed as a multi-target regression we could not find any traditional machine learning method for oversampling the data. Thus, we decided to use deep learning techniques to augment the data.

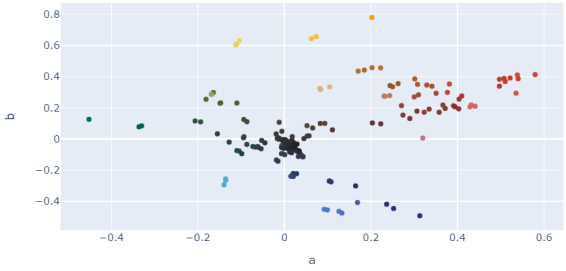


Figure 8: Plot of the train set of the WSPM dataset in L*a*b* color space.

We used the Synthetic Data Vault [13] library in order to use the state-of-the-art deep learning techniques to generate synthetic data: conditional-tabular generative adversarial networks (CT-GANS) and tabular variational auto encoders (TVAE). We tried both models, however, these models did not improve the performance. Actually, we even discovered some problems with the data. From the domain knowledge, we know that in the columns representing the ratio of the pigments, the values should only be between 0-1. However, both these models generated samples with values outside of this range. In Figures 9a and 9b, we can observe four synthetically created recipes generated with the two deep learning models. Each row represents a single recipe. The first two circles represent the pigments present in the mixture, the third one represents the target color that the generative models create ("real"), and the last circle is the predicted color from our best model (pred). Considering basic color theory and the expected outcome for the considered mixtures, it becomes clear that the results provided by the synthetically created recipes are unreliable, and, consequently, should not be used for training.

(a) CT-GAN

(b) T-VAE

Figure 9: Synthetically created recipes generated by two DL methods for the WSPM dataset. The first circles are the components. The third circle is the synthetic generated target. The final circle is the prediction made by the model.

5.2 Matcerâmica Ceramics Spectral Measurement dataset

The work done in Section 5.1 regarding the WPSM dataset provided an excellent foundation for the work to be developed on the CSM dataset. In a similar approach to the results obtained with the NTU dataset, we used the Kubelka-Munk equation as a baseline. When using the opaque readings for each component, we obtained a ΔE^*_{ab} of 26.07. Besides, when unfolding each component to use both readings a ΔE^*_{ab} of 16.45 was achieved, showing the improvements of using this unfolding strategy. Based on the gathered knowledge,

Table 4: Results of simple model with some model optimizations

batch size	Model Size	Avg Loss	% samples ΔE^*_{ab} <2	% samples ΔE^*_{ab} <5
32	5 Layers	11.36	52.38	52.38
64	5 layers	12.63	42.86	47.62

Table 5: Data-centric optimization steps results for de CSM dataset. Each row is the best result for a specific step in the optimization process.

step	Quantity Representation	Augmentation	Component Representation	Order/shuffle	Avg Loss	% samples ΔE^*_{ab} <2	% samples ΔE^*_{ab} <5
I a	0 – 1%	-	Opaque readings	-	9.564	33.33	47.62
II b	0 – 100%	-	Opaque readings	-	11.24	38.1	47.62
III c	0 – 100%	self augmentation	Opaque readings	-	10.16	47.62	47.62
IV d	0 – 1%	component augmentation	Opaque readings	-	10.85	42.86	47.62
V e	quantity	self augmentation and component augmentation	Opaque readings	-	10.41	52.38	52.38
VI f	quantity	-	Opaque readings	shuffled 3 times	6.234	57.14	71.43
VII g	quantity	-	Opaque readings	order	3.207	52.38	85.71
VIII h	0 – 1%	component augmentation	Sum	shuffled 3 times	15.01	0	4.762
IX i	0 – 100%	component augmentation	Both	self augmentation	2.35	47.62	85.71

a) Convert quantities to ratios (between 0-1).
b) Convert quantities to percentages (between 0-100%)
c) Create new recipes where only one component has a percentage of 100% (from existing recipes).
d) Create a new recipe from each component (i.e. a recipe a single component).
e) Apply simultaneously step C and D
f) Create new recipes with the order of the components shuffled (from existing recipes)
g) Order the components of each recipe by their proportions (descending order from left to right).
h) Instead of using only opaque glaze readings for each component, for each nm value of the spectral data use the weighted sum between opaque and transparent readings (according to the percentage used for that specific component).
i) Unfold each component reading into two components: with 100% opaque glaze and 100% transparent glaze. The relative amount of the respective glaze in the recipe defines the quantity percentage of each unfolded component.

we decided to only use the L*a*b* values as the target data and ΔE^*_{ab} as the loss function. Due to time constraints, we also used 1000 as the number of epochs. In the model-centric optimization phase, we assessed the impact of changing the depth of the model and batch size. To evaluate the performance, we used both $\Delta E^*_{ab} < 5$ and $\Delta E^*_{ab} < 2$, as justified in Section 4.3. Table 4 shows the two best results for the described hyperparameters using grid search.

Unfortunately, these initial results are far from the performances obtained in the WSPM dataset. One of the possible reasons might be the size of the dataset, since the CSM dataset is an order of magnitude smaller than the WPSM dataset. To overcome this limitation, data-centric optimizations were following explored to improve the models' performance. As explained in Section 4.2, we explored a wide range of data-centric optimizations. To simplify the analysis of the results, we are going to present the best results for each data augmentation step. Additionally, we will also present some interesting cases that we observed. The criteria to select the best-performing model for each optimization step was based on the average ΔE^*_{ab} obtained for the test set. Table 5 presents the results for applied data-driven optimizations, iteratively executed and compared.

To better understand these results we have to gain intuition about each one of the steps described in Section 4.2. The first two steps are fairly simple to understand: instead of dealing with quantities of the components, it is a better option to convert these quantities to percentages. As we can see, the average loss of both operations allowed us to improve the average loss on the dataset. However, the amount of samples with a ΔE^*_{ab} less than both thresholds has decreased. The following four steps (III, IV, V, VI) are trying to increase the number of recipes by leveraging the domain knowledge. From these, only step VI showed a significant improvement over the initial model-centric optimization, not only improving the ΔE^*_{ab} but also reducing the average loss in the test set for almost as half. Step VII orders the components by their quantity, the most used component is the first on the list. This step reduces in half the previous error. Until now for each component, we were using the readings made on an opaque glaze, even though in a recipe we could have a transparent glaze. The final two steps try to improve this point, by trying to create a representation of each component related to the amount of transparent and opaque glazes present in the recipe. Step H did not improve the overall metrics. The last step allowed us to obtain the best results, by reducing the average error to 2.35. This means that in the whole test set the average ΔE^*_{ab} is almost indistinguishable to the human eye.

Figure 10 shows the different colors in the test set and the prediction of the model after step IX. As we can see (recall that these colors might not be exactly how they are in real life, this is just for us to have an intuition), even though there are some discrepancies, such in idx 6 the overall predictions are almost indistinguishable to the human eye. Spite the fact this dataset is considerably smaller than the WSPM, we showed the feasibility of using deep learning algorithms to improve the color mixture predictions.

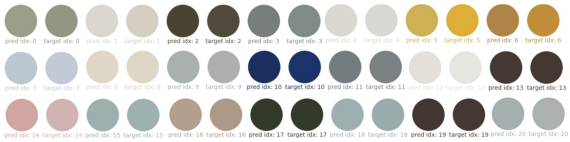


Figure 10: Predictions of the best model on the CSM test set. For each recipe, we have two circles, on the left is the prediction, and on the right, we have the target.

6 CONCLUSION

This paper explored the usage of data-centric Deep Learning techniques to improve color mixture predictions in ceramic glazes. While the usage of deep learning techniques is starting to emerge in similar areas of applications, to the best of our knowledge, the usage of data-centric approaches in this type of problem, either automated of domain-driven, is firstly introduced in this work.

The exploratory work performed on the WPSM public dataset allowed to validate the relevance of the proposed data-centric steps, which brought significant performance improvements. The knowledge gained on this initial study was then transferred to the target dataset Matcerâmica Ceramics Spectral Measurement dataset, which was specifically created in the ambit of this work. Even though the areas of application of both datasets are different, one is in paper and watercolors and the other in ceramics, there are a lot of similarities (recipe structure and the type of the data). In the CSM dataset, we showed that even though the size is not comparable, after the data-centric optimizations we achieved promising results for automated color mixture prediction.

Regarding future work, the size of the CSM dataset is currently being continuously increased, which will allow us to keep using data-centric optimizations to improve the overall results. We also plan to use these models in a real-world environment, by integrating them into a computer-aided tool that can be deployed in ceramics color laboratories. While using these models as an assisting tool, we aim to perform a direct comparative analysis between the traditional method of physically creating color mixtures and use these models as a digital tool to predict color mixture results. Particularly, this analysis will help us quantify the efficiency gains regarding aspects such as time spent or the amount of energy saved by removing the need of cooking unnecessary ceramic pieces during color adjustment processes.

ACKNOWLEDGMENTS

This work was financially supported by European Regional Development Fund (ERDF) in the frame of Norte 2020 (Programa Operacional Regional do Norte), through the project SmartColor4Ceramics – Smart Color Replication for Ceramics, with reference POCI-01-0247-FEDER-069890.

REFERENCES

[1] JM Bishop, MJ Bushnell, A Usher, and S Westland. 1991. Neural networks in the colour industry. In *Applications of Artificial Intelligence in Engineering VI*. Springer, 423–434.

[2] JM Bishop, MJ Bushnell, and S Westland. 1991. Application of neural networks to computer recipe prediction. *Color Research & Application* 16, 1 (1991), 3–9.

[3] Edgar Andrés Chavarriaga Miranda, Nelson Alonso Correa Rojas, Juan Fernando Montoya Carvajal, and Oscar Jaime Restrepo Baena. 2015. Color prediction in ceramic enamels using the kubelka munk model. *Tecciencia* 10, 19 (2015), 1–5.

[4] Andrew Chen, Andy Chow, Aaron Davidson, Arjun DCunha, Ali Ghodsi, Sue Ann Hong, Andy Konwinski, Clemens Mewald, Siddharth Murching, Tomas Nykodym, Paul Ogilvie, Mani Parkhe, Avesh Singh, Fen Xie, Matei Zaharia, Richard Zang, Juntai Zheng, and Corey Zumar. 2020. Developments in MLflow: A System to Accelerate the Machine Learning Lifecycle. In *Proceedings of the Fourth International Workshop on Data Management for End-to-End Machine Learning* (Portland, OR, USA) *(DEEM'20)*. Association for Computing Machinery, New York, NY, USA, Article 5, 4 pages. https://doi.org/10.1145/3399579.3399867

[5] Mei-Yun Chen, Ya-Bo Huang, Sheng-Ping Chang, and Ming Ouhyoung. 2019. Prediction Model for Semitransparent Watercolor Pigment Mixtures Using Deep Learning with a Dataset of Transmittance and Reflectance. *arXiv preprint arXiv:1904.00275* (2019).

[6] Rocco Furferi and Monica Carfagni. 2010. Colour mixing modelling and simulation: optimization of colour recipe for carded fibres. *Modelling and Simulation in Engineering* 2010 (2010).

[7] Rocco Furferi, Lapo Governi, and Yary Volpe. 2016. Color matching of fabric blends: Hybrid Kubelka-Munk+ artificial neural network based method. *Journal of Electronic Imaging* 25, 6 (2016), 061402.

[8] Caroline Hemingray and Stephen Westland. 2016. A novel approach to using neural networks to predict the colour of fibre blends. *Coloration Technology* 132, 4 (2016), 297–303.

[9] Ya-Bo Huang, Mei-Yun Chen, and Ming Ouhyoung. 2018. Perceptual-based CNN model for watercolor mixing prediction. In *ACM SIGGRAPH 2018 Posters*. 1–2.

[10] Malathy Jawahar, Chandra Babu Narasimhan Kannan, and Mehta Kondamudi Manobhai. 2015. Artificial neural networks for colour prediction in leather dyeing on the basis of a tristimulus system. *Coloration Technology* 131, 1 (2015), 48–57.

[11] Paul Kubelka and Franz Munk. 1931. An article on optics of paint layers. *Z. Tech. Phys* 12, 593-601 (1931), 259–274.

[12] M Ronnier Luo, Guihua Cui, and Bryan Rigg. 2001. The development of the CIE 2000 colour-difference formula: CIEDE2000. *Color Research & Application: Endorsed by Inter-Society Color Council, The Colour Group (Great Britain), Canadian Society for Color, Color Science Association of Japan, Dutch Society for the Study of Color, The Swedish Colour Centre Foundation, Colour Society of Australia, Centre Français de la Couleur* 26, 5 (2001), 340–350.

[13] Neha Patki, Roy Wedge, and Kalyan Veeramachaneni. 2016. The synthetic data vault. In *2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*. IEEE, 399–410.

[14] LM Schabbach, F Bondioli, AM Ferrari, CO Petter, and MC Fredel. 2009. Colour in ceramic glazes: Efficiency of the Kubelka–Munk model in glazes with a black pigment and opacifier. *Journal of the European ceramic society* 29, 13 (2009), 2685–2690.

[15] Azadeh Asadi Shahmirzadi, Vahid Babaei, and Hans-Peter Seidel. 2020. A Multispectral Dataset of Oil and Watercolor Paints. *Electronic Imaging* 2020, 5 (2020), 107–1.

[16] Yi Wang, Qinxin Zhao, Jian Chen, Weiwei Wang, Suihuai Yu, and Xiaoyan Yang. 2022. Color Design Decisions for Ceramic Products Based on Quantification of Perceptual Characteristics. *Sensors* 22, 14 (2022). https://doi.org/10.3390/s22145415