IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 10 December 2023, accepted 7 February 2024, date of publication 14 February 2024, date of current version 26 February 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3366491

RESEARCH ARTICLE

Identification of Algal Blooms in Lakes in the Baltic States Using Sentinel-2 Data and Artificial Neural Networks

DALIA GRENDAITĖ[1] AND LINAS PETKEVIČIUS[2], (Member, IEEE)
[1]Institute of Geosciences, Vilnius University, 03101 Vilnius, Lithuania
[2]Institute of Computer Science, Vilnius University, 08303 Vilnius, Lithuania

Corresponding author: Dalia Grendaitė (dalia.grendaite@chgf.vu.lt)

This work was supported in part by the European Union through the Research Council of Lithuania (LMTLT) under Project S-MIP-23-44 and in part by the Vilnius University Science Promotion Fund for supporting initial stage of research which extended to LMTLT project and this publication under Grant MSF-JM7/2021.

ABSTRACT Algal blooms are a common problem in inland waters, which raise growing awareness on monitoring lakes' conditions. The on site monitoring is expensive and requires large human resources efforts. This work proposes remote monitoring techniques using satellite images and machine learning algorithms to predict chlorophyll α concentration in water bodies and identify algal blooms. The training and test dataset used in this study includes diverse range of lakes in Baltic countries. The lake spectral features obtained from Sentinel-2 satellite images are used as predictors for proposed deep neural network models. The prediction of chlorophyll α concentration with MAE 7.97 mg/m^3 and bloom vs. non-bloom classification with 71.6 % accuracy was achieved. The use of Bèzier curves for smoothing the point-wise prediction is proposed for identification of algal bloom characteristics: the bloom start date, end date, and duration. The results showed lake type impact on the blooming time. The experimental data and code are released with paper.

INDEX TERMS Satellite image processing, chlorophyll α prediction, deep neural networks, remote sensing, Bèzier curves.

I. INTRODUCTION

Algal and cyanobacterial blooms are a natural phenomenon that are caused by sudden proliferation of algal and cyanobacterial biomass. These organisms take up carbon dioxide and release oxygen during photosynthesis. Additionally, they are food to zooplankton and fish. However, human pressures such as increased discharge of pollution and alterations of water bodies [13] cause more intensive algal and cyanobacterial blooms. These blooms result in low water transparency and limit light penetration to deeper water layers, thus ecosystem structure changes and benthic plants may disappear. Increased turbidity also causes changes in fish communities. Algal blooms decrease oxygen levels and cause fish suffocation. In addition, a large part of intense blooms globally are toxic [14], thus they are called harmful algal blooms (HABs). Toxins produced by some algae and cyanobacteria are not only dangerous to fish and other organisms in water but also to people's health. The various toxins produced by algae and cyanobacteria can lead to allergies, increased risk of illnesses, and death [17]. Algal blooms have been reported in various locations across the world - in tropical [40], temperate regions [4], [6], and since 2000s they have been often observed in the northern latitudes in the arctic lakes [21].

Climate change brings more challenges to ecosystems and water use to people as increased temperatures, longer stratification periods, decrease of ice cover will lead to more suitable conditions for algal and cyanobacterial blooms across the world [17]. In addition, increasing trends in algal bloom frequency have been observed in 467 lakes with an area larger than 1 km^2 over almost 40 year period; however anthropogenic factors had a higher impact than climatic drivers to algal bloom intensification [8]. Thus it is important to observe algal blooms, their frequency, onset, length, intensity, and extent [45]. Complementing in situ data

The associate editor coordinating the review of this manuscript and approving it for publication was Chuan Li.

with satellite data can also improve the reporting of algal blooms by member states to European Union as now they are under-reported [45].

Cyanobacterial blooms occur in a number of lakes, in a large Kaunas reservoir and cyanobacteria hyperblooms in the Curonian lagoon [56]. Cyanobacterial orders produce harmful toxins [20]. A study based on satellite data and 226 lakes in Lithuania revealed that 42 lakes are frequently affected by algal blooms [12]. In Estonia 15% of lakes were reported to be affected by cyanobacterial blooms [47]. However, a larger scale research that would include all the larger lakes in these countries is missing.

Algal blooms have been observed from space from 1970s when first satellites were launched. Algal blooms were observed in oceans as due to their specific optical characteristics they change water colour. Current satellite sensors allow to observe even small water bodies and spot water circulation patterns that are enhanced by algal biomass distribution in larger lakes and seas.

Algal blooms can be described by many parameters, including the total biomass, the biomass of different orders, the number of cells in mL; however, from satellite data the closest parameter that we can use to define the algal bloom is the concentration of chlorophyll α – the pigment common for all algae and cyanobacteria. The concentration of 10 mg/m^3 is considered as increased algal biomass and posing a risk [44], it is used as alarm level 1 [39], and is also recognised as a lower range of concentrations that can be detected using some satellite data-based algorithms [1].

There are two approaches for deriving water quality parameters from satellite data – the model-based approach that model remote sensing reflectance in terms of inherent optical properties of water using radiative transfer modelling and the second approach – empirical approach finds a relationship, often using linear regression, between in situ measured water quality parameter data and remotely sensed data [36]. Recently a lot of empirical site and sensor-specific algorithms have been developed [30], [36], [41]. In the case of waters in which one type of optically active substance is prevalent, such as chlorophyll α in the Case I – ocean waters [38], the algorithms that use shorter wavelengths (blue and green) work fine [43]. However, in the case of presence of other optically active substances in water (Case II waters), such as, suspended matter and coloured dissolved organic matter that do not covary with chlorophyll α, simple algorithms may fail to work well due to the interfering signal of these substances [25]. In such cases it is better to group the similar waters together and compose algorithms for distinct water types [37], [53]. In addition, more complex algorithms, such as, machine learning-based algorithms, that can extract relevant information from multi-dimensional data, provide better results [48], especially when used with hyperspectral data [22]. In addition, machine learning algorithms can account for complex interactions among variables, thus machine learning algorithms, such as, boosted regression trees and random forest algorithms showed better predictive skill of chlorophyll α concentration than multiple linear regression in American lakes [32]. However, models based on artificial neural networks often show the best accuracy in satellite data analysis [27], including the retrieval of water quality parameters from remotely sensed data [29], [33], [34], [50], [59]. The feed forward/backpropagation artificial neural networks have been used for retrieving concentrations of chlorophyll α [49], suspended matter, yellow substances [54], turbidity [29], and inherent optical properties [18]. Moreover, a combination of artificial neural networks has been used to perform atmospheric correction, retrieve inherent optical, and derive water constituents [2], [5].

Recently, there have been developments in the field of deep learning and new artificial neural network configurations may increase the accuracy of retrieval of concentrations of optically active substances even more. Thus the aim of this study was to use advanced deep learning models to separate the blooming waters from non-blooming waters using only satellite data and then quantitatively estimate the intensity of blooms using the derived chlorophyll α concentration from satellite data. In addition, to cover a wide variety of water bodies, the dataset used in this study is from three Baltic states: Lithuania, Latvia, and Estonia and comprises data from 742 monitoring sites across these countries. The main contributions of this paper are:

a) Developed novel neural network model was able to achieve state of the art classification of bloom accuracy 71.6%, and 7.97 mean absolute error for chlorophyll α concentration;

b) The proposed methodology to identify blooming start and end times by smoothing point-wise chlorophyll α predictions using Bèzier curves.

The paper is organized as follows: first, results of the created algorithms for classification of bloom presence, secondly, the discussion of results are presented, thirdly design and setup of our experiment environment are presented and conclusions close the article. The used mathematical methods are described in Appendix.

II. METHODS

A. STUDY AREA AND IN SITU DATASET

The Baltic states according to Koppen climate classification are in Dfb zone that is characterised by humid continental climate with warm summers. The annual mean daily temperature in Lithuanian continental meteorological stations is 7.9 C, in Latvian – 7.8 C, and in Estonian continental stations it is 8.2 C, while in maritime stations the annual temperature is 5.8 C, in Latvian – 6.6 C, and in Estonian maritime stations it is 5.3 C [19]. The countries lie in the western part of East European plain. Diffuse pollution from agricultural fields is the main pressure on surface water in the region.

The lakes and ponds included in this study are a part of the national monitoring programmes in the Baltic countries. There are 357 lakes and ponds in Lithuania, 276 lakes and ponds in Latvia, and 80 lakes and ponds with 109 monitoring sites in Estonia (Figure 1), in total 742 monitoring sites. Most

of the lakes are larger than 0.5 km^2. We collected in situ data from the years of 2017-2021 of chlorophyll α concentration and water transparency from national environmental agencies of Lithuania, Latvia, and Estonia (Table 1).

The data analysis was also carried out considering lake type that is defined based on mean and maximum depth of a lake. The lake typology is defined by Lithuanian Ministry of Environment [7]. The type 1 lakes are shallow (average depth < 3 m or average depth > 3 m and maximum depth < 11 m), the lakes of type 2 are medium deep (average depth > 3 m and maximum depth 11–30 m), and type 3 are deep lakes (maximum depth > 30 m).

For deeper analysis we selected 105 measurement sites that were in 99 lakes, as these measurement sites had six to 30 in situ observations during the years of 2017-2021. The 68 lakes were in Lithuania, 14 in Latvia, and 23 measurement sites were in Estonia, from which 7 were in Lake Peipsi. There have been 62 shallow lakes (59%), 35 medium deep (33%), and 8 deep lakes (8%).

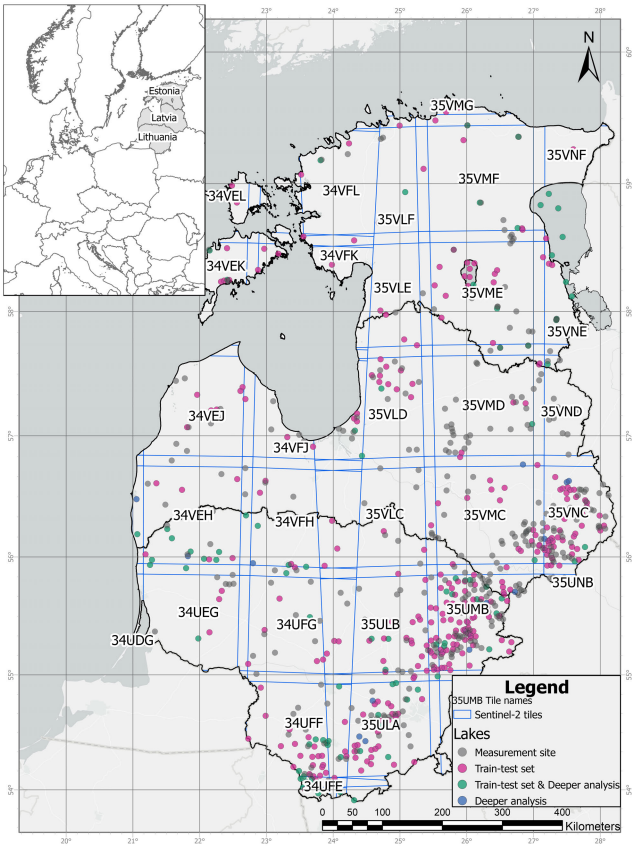

FIGURE 1. The location of monitoring sites (grey), points that were used for model training and testing (pink), points that were both in training-test set and for deeper analysis (green), and points that were not in training-test set but were used for deeper analysis (blue) in Lithuania, Latvia, and Estonia.

B. SATELLITE DATASET

We used optical Sentinel-2 MultiSpectral Imager (MSI) data. We extracted pixels around monitoring sites using Google Earth Engine python API [10]. Sentinel-2 MSI has 13 bands in visible (443-665 nm, B1-B4), near-infrared (705-865 nm, B5-B8A), and short wave infrared range (940-2190 nm, B9-B12) with spatial resolution of 10, 20, and 60 m. The dataset was then filtered retaining the pixels that were flagged as water by scene classification algorithm used in Sen2Cor. Also, we removed all the spectra where surface reflectance was higher than 0.0215 in the shortwave 1610 nm (B11) band and the spectra where reflectance in the blue band (B2) where lower than 0.001 to avoid the spectra that where not water but were missed by the Sen2Cor algorithm.

Our training-test dataset comprised of 1346 corresponding in situ and satellite observations across 477 monitoring sites in the Baltic lakes and ponds. The chlorophyll α concentration range in the dataset was 0.2-167.2 mg/m^3 (mean = 15.6 mg/m^3, sd = 20.3 mg/m^3).The observations were considered corresponding when there was no more than three days time (backwards or forward) difference between in situ measurement and satellite acquisition. The 44% of observations were affected by light to strong algal blooms More than half of observations included in train-test dataset were in Lithuania (55%), while a similar number of observations were from Latvian (20%) and Estonian lakes (25%) (Table(1). Algal bloom was defined when chlorophyll α concentration was 10 mg/m^3 or higher.

C. THE CALCULATION OF BLOOM PARAMETERS

We analysed data from 105 measurement sites that were selected based on in situ data availability, that in turn show correspondence model vs. in situ measured values. The 105 measurement sites were in 99 lakes and had 6 to 30 in situ measurements during the five-year period (2017–2021). For these 105 measurement sites we analysed the duration of blooms for years 2017 to 2021. For year 2017 we had no data for 9 lakes. Also, since the year 2017 was a rather cloudy year, thus there were fewer observations available and we used only those cases where there were at least five satellite data points in a year, thus, additionally 37 lake data were removed as insufficient.

We calculated bloom characteristics from the modeled and smoothed data: the start date, end date, duration, mean chlorophyll α concentration, and maximum chlorophyll α concentration. The bloom start and end dates were determined when chlorophyll α concentration rose higher than 10 mg/m^3, usually in spring or early summer time, and when it went down and dropped below 10 mg/m^3 in late summer or autumn time. In some occasions there were a few such periods during a year. In these cases the duration of bloom in a lake in that year was calculated as the sum of blooming episodes.

D. MACHINE LEARNING METHODS

Machine learning techniques and algorithms are capable of capturing trends in data and using them for estimation and decision making tasks [9]. In this research we will use supervised learning – algorithms which use data sets of

TABLE 1. Characteristics of train-test dataset by country.

Country	Bloom label	Number of lake observations	Mean ± standard deviation of chlorophyll α concentration, mg/m^3	Mean ± standard deviation of of transparency, m
Lithuania	0	447	4.5 ± 2.4	3.9 ± 2.0
	1	290	31.3 ± 26.8	1.6 ± 1.0
Latvia	0	152	4.6 ± 2.6	2.0 ± 1.0
	1	121	27.6 ± 22.9	1.2 ± 0.5
Estonia	0	155	5.0 ± 2.6	2.3 ± 1.2
	1	181	27.9 ± 20.2	1.0 ± 0.6

associated pairs of input features and output values target. Depending on the output type of the algorithm they can be further divided into classification and regression tasks. In our binary classification task the given feature vectors are assigned a discrete enumerated value zero or one. Formally, supervised learning can be formulated as search of operator f, where $y = f(x|\theta) : \mathbb{R}^{d_x} \rightarrow \mathbb{R}^{d_y}$, where x – input values with dimension of d_x, y – output values of dimension of d_y and $\theta \subset \Theta \in \mathbb{R}^{d_\theta}$ – unknown parameters of operator. In deep learning the unknown parameters estimation (learning) is conducted by optimizing some objective function $\mathcal{L}$, with respect to unknown parameters, which evaluate the difference between ground truth target values and values generated by the algorithm. In our experimental setup the Cross-entropy loss function is used for the binary-class problem and mean square error for regression problem [9].

The classical machine learning models were used for creation of benchmark models (See Section II-H). The custom deep neural network models were created as our experiments demonstrated they performed significantly better than baseline models. We constructed custom neural networks using fully connected, dropout, batch normalization, feature selection, attention mechanism, and wide and deep layers. All details about those neural networks transformations are described in the Supplementary Material.

E. DATA PREPROCESSING FOR MACHINE LEARNING MODELS

The dataset was divided into three parts (training, validation and testing) based on different lakes to avoid data leakage problem. The 14 variables were used by ML/deep learning models as input (independent) variables and the chlorophyll α as output (dependent) variable to be inferred by the models. For machine learning the package `pycaret` was used to benchmark classical ML models. In experimentation models removed multicollinearity with threshold $T = 0.4$. K-fold with 10 folds were used to run benchmark models. Equivalently the dataset was used in deep learning models. No specific feature engineering, data preprocessing or normalization was applied.

F. CLASSIFICATION MODEL PERFORMANCE METRICS

In the study we used the following performance measures for evaluation of accuracy of bloom classification models [9]: accuracy, precision, recall, and F1-score. All of these statistics can be calculated from the classification table which pivot predicted and actual observed conditions. We denote shorter notations as True positive (TP), False negative (FN), False positive (FP) and True negative (TN). Classification accuracy shows how accurately the model predicts investigated bloom/non-bloom classification Accuracy $= (\mathrm{TP} + \mathrm{TN})/(\mathrm{TP} + \mathrm{FN} + \mathrm{FP} + \mathrm{TN})$. Precision and recall are also used as classification accuracy metrics. Precision shows the ratio of TP between all positively predicted samples, while recall is the ratio of TP between all truly positive samples Precision $= \mathrm{TP}/(\mathrm{TP} + \mathrm{FP})$, Recall $= \mathrm{TP}/(\mathrm{TP} + \mathrm{FN})$. One more measure F1-score, is a so-called weighted mean of precision and recall where, F1-score is treated as harmonic mean of the precision and recall and is expressed as F1-score $= 2\mathrm{TP}/(2\mathrm{TP} + \mathrm{FP} + \mathrm{FN})$.

G. REGRESSION MODEL PERFORMANCE METRICS

In the study for regression model evaluation we used the following metrics: mean absolute error (MAE), mean squared error (MSE), root mean squared error (RMSE) and determination coefficient (R^2). All of these metrics can be calculated using the ground truth and predicted values of the target variable. The MAE is the average of the absolute differences between the predicted and actual values. The MSE is the average of the squared differences between the predicted and actual values. The RMSE is the square root of the MSE. The R^2 is the proportion of the variance in the target variable that is predictable from the input variables. The R^2 is a value between zero and one, where zero means that a model explains none of the variability of the response data around its mean, and one means that the model explains all the variability of the response data around its mean. The R^2 is also known as the coefficient of determination.

H. ARTIFICIAL NEURAL NETWORK MODELS

1) A MODEL FOR ALGAL BLOOM IDENTIFICATION

We labeled observations as 'bloom' (class value 1) when the chlorophyll α concentration was equal to or higher than 10 mg/m^3 and lower concentrations as 'non-bloom' (class value 0). For binary classification further we investigate the various machine learning classifiers. Twenty nine models were developed for distinguishing bloom and non-bloom lakes from satellite aggregated features. The experiment was designed to have some benchmark machine learning models and later create custom new neural networks to improve the predictions. For benchmark models `pycaret` package was used. Package contains the list of 15 machine

learning models (Gradient Boosting Classifier, Random Forest Classifier, Light Gradient Boosting Machine, Ada Boost Classifier, Extra Trees Classifier, Extreme Gradient Boosting, K Neighbors Classifier, Linear Discriminant Analysis, Ridge Classifier, Logistic Regression, Decision Tree Classifier, Naive Bayes, SVM – Linear Kernel, Dummy Classifier, Quadratic Discriminant Analysis), which were used as benchmark for the experiment.

The custom neural network models were created by experimenting with various new layers, and optimizing the proposed architectures. The dropout of 0.2 was used in many layers to ensure robustness of solution [9]. The 14 models were proposed, out of which some outperformed the machine learning benchmark models.

Training and test data sets contained 70% and 30% of total separate lakes to avoid data leakage of having some measurements on the same lake. The 477 lakes were separated to 333 for model creation (training) and independent 144 unseen for model lakes were left for independent testing. Data sets then contained 909 and 437 lake measurements in subset splits, respectively.

We used surface reflectance of the 490-865 nm wavelength bands (B2-B8A), band ratios: R705/R665, R560/R490, R560/R665, R560/R705, band difference (BD) that is based on the difference between the R705 and the average of R665 and R740 [55], the empirical equation derived for Lithuanian lakes that use R705 and R665 (Eq_diff) [11], and Apparent Visible Wavelength (AVW) [57], that represents colour parameter from visible wavelength bands (B2 (R490), B3 (R560), and B4 (R665)).

In this study, multiple configurations of neural networks architectures were considered, the default classification threshold $T = 0.5$ was used in model creation. Experimentally, the best classifying models for bloom classification were identified. The architecture of those models are presented in Table 2. For all models, training was carried out by monitoring test loss with epoch size up to 50, with early stopping parameter of 20 epochs and batch size of 16. The binary cross-entropy loss function and Adam [9], [24] optimizer with learning rate `0.001` was used for neural network parameters estimation.

2) A MODEL FOR CHLOROPHYLL α RETRIEVAL

We used chlorophyll α observations as regression targets in regression models. As in classification experiments we bench-marked the ML models. In the experiment setup using `Pycaret` package 19 regression models (Passive Aggressive Regressor, Huber Regressor, Orthogonal Matching Pursuit, Lasso Regression, Elastic Net, Bayesian Ridge, Ridge Regression, Dummy Regressor, Least Angle Regression, Linear Regression, Gradient Boosting Regressor, CatBoost Regressor, Extra Trees Regressor, Random Forest Regressor, K Neighbors Regressor, Light Gradient Boosting Machine, Extreme Gradient Boosting, Decision Tree Regressor, AdaBoost Regressor) were tested. The data splitting and experimentation setup, was equivalent as in the classification experiments. Since we created multiple deep neural network regression models, we investigated multiple loss functions. What we found and used in the final experimentation was the Huber loss function [16]. The Huber loss function is less sensitive to outliers in data than the Mean Squared Error loss function. The Huber loss function is defined as:

$$L_\delta(y, f(x)) = \begin{cases} \frac{1}{2}(y - f(x))^2 & \text{if } |y - f(x)| \le \delta \\ \delta(|y - f(x)| - \frac{1}{2}\delta) & \text{otherwise} \end{cases}$$

where y is the true value and $f(x)$ is the predicted value. The parameter $\delta = 10$ is the threshold value. The selection of Huber loss penalised lower errors than Mean Squared Error loss function. In practice this allowed to have more precise predictions for the chlorophyll α on low concentrations. In practice we witnessed that to precisely identify level-1 alarm was more important than precisely predict very large concentrations of chlorophyll α.

3) PREDICTIONS SMOOTHING USING BÈZIER CURVES

The results of classification model were used to filter chlorophyll α concentrations obtained with regression model in order to synchronise results – when classification model showed non-bloom conditions, chlorophyll α concentration should have not been higher than 10 mg/m^3 and when blooming conditions were determined by classification model, the higher chlorophyll α concentrations than 10 mg/m^3 were expected. The 9.6% of data were filtered out in this way before further analysis.

The machine learning predictions are often noisy due to the fact that we make predictions on independent satellite data. Due to clouds and atmospheric conditions, the model predictions may be noisy. However, we can improve the robustness of predictions by smoothing the machine learning predictions. The temporal data during the year does not change rapidly, in such the smoothness of the prediction can improve the robustness of estimates. In our study we propose to use Bèzier curves [15], [51]. Bèzier curves are a family of parametric curves used in computer graphics and related fields. The curve is defined by a set of selected points and a parameter τ that varies between 0 and 1.

The Bèzier curve is a linear combination of the selected points, where the weights are defined by the Bernstein polynomials. Let's assume that we have a set of points $w_k = (t_k, y_k)$, where y – represents prediction, t – yearly time, $k = 0, 1, \cdots, m$. Then the m-th order Bèzier curve is defined as $f(\tau)$: where $\tau \in [0, 1]$ using $m + 1$ points,

$$\begin{aligned} f(\tau) &= w_0 b_m^0(\tau) + w_1 b_m^1(\tau) + \cdots + w_m b_m^m(\tau) \\ &= \sum_{i=0}^{m} w_i \cdot b_m^i(\tau), \end{aligned}$$

where w_i prediction point, and $b_m^i(\tau) = \binom{m}{i}\tau^i \cdot (1-\tau)^{m-i}$ Bernstein base. The derivative of Bèzier curve $\frac{df(\tau)}{d\tau}$ are

TABLE 2. Configurations of classification models.

Model	Architecture/hyper-parameters
Baseline	FL-D-FC(4*H)-BN-Relu-D-FC(H)-BN-Relu-D-FC(1)
Residual DNN	FL-BN-FC(H)-Relu-BN-Res(1,2,1)-FL-FC(1)
Residual DNN + MHA	FL-BN-MHA(FC(H), 4)-Res(2)-FL-FC(1)
Residual DNN + 2xMHA	FL-BN-MHA(FC(H), 4)-Res(1,2,1)-FL-MHA(FC(8), 2)-FC(1)
Gated Residual1 + VS + FC	VS(H)-FC(1)
Gated Residual2 + VS	VS(H)-BN-FC(16)-FC(1)
Gated Residual3 + VS + FC	VS(H)-FC(1)
Gated Residual4 + VS	VS(H)-BN-FC(64)-FC(1)
Gated Residual5 + MHA	VS(H)-BN-FC(H)-MHA(FC(H), 8)-FC(1)
DNN + MHA	FL-MHA(FC(H), 8)-FL-FC(1)
Wide Deep DNN	FL-BN-Wide-MHA(FC(32), 8)-Deep(4*H,16)-MHA(FC(H), 8)-FL–MHA(FC(16), 8)-FC(1)
Wide Deep DNN + MHA	FL-BN-Wide-Deep(4*H,H)-FC(1)
Cross Deep DNN	FL-Cross(128,16)-Deep(4*H,H)-FL-FC(1)
Cross Deep DNN + MHA	FL-MHA(FC(32), 8)-Cross(4*H,H)-Deep(4*H,H)-FL- MHA(FC(16), 8) -FC(1)
CNN	Conv-BN-Relu-MaxPool-FL-FC(H)-BN-Relu-D-FC(1)

The model architectural representations. FL – stands for Flattening, D – for Dropout (p=0.2), BN – for batch normalization, Res – for residual block, MHA – Multi-head attention, VS – continuous variable selection layer, Wide – wide connection, Deep – deep connection, Conv – convolutional connection, neural network layers and Relu is non-linear point-wise activation. The H – is hidden state parameter for classification models $H = 32$ in regression experiment setup $H = \{4, 8, 16, 32, 64, 128\}$.

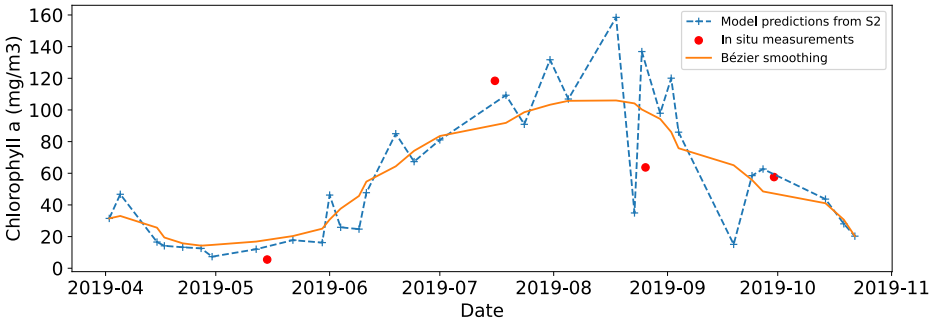

FIGURE 2. Bèzier smoothing for Lake Lateženis (LT). The dotted blue line represents machine learning predictions, while the solid brown line represents Bèzier curve. In situ measurements are shown as red dots.

intersect points $p_i^{(1)} = m \cdot (p_{i+1} - p_i)$. An example of Bèzier curve in our experiments is shown in Figure 2.

The main benefit which prediction smoothing enables is to investigate the data within the region of interest. In situ measurements are rare and sparse due to diversified data collection (Figure 3). On the other hand, the smoothing enables to interpolate predictions' values on daily level. This enables to discover new patterns like, start and end of blooming timings, dependencies on chlorophyll α changes.

III. RESULTS

A. MODEL FOR SEPARATION OF BLOOM FROM NON-BLOOM CONDITIONS

The extensive investigation of multiple deep learning models was carried out. Each deep neural network (DNN) model was trained, and saved for independent performance evaluation on test set. The best performing Gated Residual Network with Variable Selection and Multi-Head attention layers model $M_{GRN,VS,MHA}$ was able to classify 84% of bloom samples and 63% of non-bloom samples of the test set (Table 3). The mean chlorophyll α concentration of the true positives (TP, blooming samples) was 31.4 mg/m^3 (sd = 24.3 mg/m^3), while of the false positives it was 16.6 mg/m^3 (sd = 6.8 mg/m^3), thus the blooming samples that were further from class border were classified correctly. The mean transparency of the FP samples was higher (2.5 m, sd = 1.2 m) than that of the TP (1.2 m, sd = 0.7 m). In some cases features' values of the FP were more similar to the values of the FN than to the TP, thus, they were misclassified. Feature overlap due to spectral similarities and strict border set between the bloom and non-bloom waters caused misclassification of 27% of samples. For example, the mean value of R705/R665 of the FP (0.87) was more similar to the value of the TN (0.94), rather than to the value of the TP (1.17) (Table 4). Usually the R705/R665 ratio is positive when water is blooming due to chlorophyll α absorption at the wavelengths around 665 nm and due to some backscattering at the wavelengths around 705 nm. Our model was able to classify a part of cases that did not have such a clear spectral peak at the wavelengths around 705 nm, thus our model is advantageous over simple band or band ratio models. Similarly, a part of cases where a clear spectral peak was observed in samples which were labelled as non-bloom

TABLE 3. The classification of bloom from non-bloom conditions models performance metrics on the test set.

Model	Accuracy	AUC	Recall	Prec.	F1	Accuracy (train)	Number of params
Gated Residual5 + MHA	0.716	**0.805**	0.622	0.763	0.713	0.760	797103
Gated Residual1 + VS + FC	0.709	0.798	0.599	0.765	0.706	0.751	1439
Gated Residual2 + VS	**0.723**	0.798	0.682	0.740	**0.723**	0.774	14975
Gated Residual3 + VS + FC	0.698	0.794	0.512	0.810	0.687	0.722	195599
Cross Deep DNN	0.707	0.772	0.691	0.711	0.707	0.723	5067
Gated Residual4 + VS	0.691	0.770	**0.783**	0.659	0.689	0.714	197903
Wide Deep DNN	0.664	0.754	0.406	**0.830**	0.639	0.696	4647
Wide Deep DNN + MHA	0.652	0.753	0.442	0.756	0.636	0.688	65129
Residual DNN + 2xMHA	0.641	0.736	0.387	0.778	0.615	0.690	142273
Baseline	0.581	0.735	0.184	0.870	0.502	0.652	6721
DNN + MHA	0.670	0.732	0.548	0.721	0.665	0.714	9153
Cross Deep DNN + MHA	0.673	0.732	0.636	0.683	0.672	0.711	211905
Residual DNN + MHA	0.670	0.725	0.535	0.730	0.664	0.737	71193
Logistic Regression	0.673	0.720	0.521	0.743	0.612	0.707	-
CatBoost Classifier	0.659	0.709	0.521	0.715	0.603	0.843	-
Gradient Boosting Classifier	0.641	0.703	0.493	0.695	0.577	0.834	-
K Neighbors Classifier	0.645	0.694	0.530	0.684	0.597	0.767	-
Ada Boost Classifier	0.622	0.688	0.516	0.651	0.576	0.765	-
SVM - Linear Kernel	0.675	0.675	0.659	0.678	0.668	0.698	-
Random Forest Classifier	0.638	0.667	0.479	0.698	0.568	1.000	-
CNN	0.497	0.664	1.000	0.497	0.332	0.413	79809
Light Gradient Boosting Machine	0.636	0.655	0.502	0.681	0.578	0.962	-
Extra Trees Classifier	0.613	0.650	0.475	0.652	0.549	1.000	-
Extreme Gradient Boosting	0.620	0.647	0.470	0.667	0.551	0.997	-
Residual DNN	0.595	0.647	0.475	0.620	0.589	0.589	36665
Linear Discriminant Analysis	0.563	0.644	0.272	0.641	0.382	0.595	-
Naive Bayes	0.563	0.626	0.917	0.535	0.676	0.518	-
Decision Tree Classifier	0.586	0.585	0.461	0.610	0.525	1.000	-
Ridge Classifier	0.554	0.552	0.254	0.625	0.361	0.583	-
Dummy Classifier	0.503	0.500	0.000	0.000	0.000	0.588	-

due to low chlorophyll α concentration, were classified as blooming (FN). These cases might show monitoring sites that are not representative of those lakes as they miss the blooming conditions or these lakes are characterised by low coloured organic dissolved matter (CDOM) content and thus the CDOM does not mask the signal of the relatively low chlorophyll α concentration and it is clearly pronounced in the spectra.

The $M_{GRN,VS,MHA}$ was selected out of top tier based on Area under the curve (AUC) statistics (Table 3). The selection of best model could also be considered based on F1 or accuracy statistics; however since we do not have imbalanced data classes the AUC covered both type of errors minimization. While model have large number of unknown model parameters, to be robust models were trained using small batch sizes ($B = 16$) and dropout ($p = 0.2$). Also, we noticed that some of the classical algorithms like Random Forest, Extra Trees Classifier while demonstrating similar results were over-fitted on training set (100%) thus, they would not generalize well over large amount of new unseen data.

B. MODEL FOR CHLOROPHYLL α CONCENTRATION RETRIEVAL

We used mean absolute error (MAE) and mean mean squared error (MSE) as quality metrics to analyze the performance of ML models to predict the chlorophyll α concentration. The wide investigation of ML models for regression problem by estimating the concrete value of chlorophyll α was delivered. The best performing model for regression Residual deep neural network with Multi-Head Attention model $M_{R,MHA}$ have achieved MAE = 7.97 mg/m^3 (Table 5). The results show that the best model by selected metrics did not always point to the same ML model as it was in the case of classification. This is due to the fact that the regression problem is more complex and models are more sensitive to the data. The low values of determination coefficient including the best 0.55 shows high variation within the data. The widest range of errors have appeared in non-blooming shallow and medium depth lakes (Table 6). In bloom conditions mean error was higher than in non-bloom conditions; however, mean absolute percentage error (MAPE) was much lower during blooming, therefore, the model was able to predict the higher chlorophyll α concentrations much better than the lower ones (during non-bloom conditions) (Table 6).

The best performing model $M_{R,MHA}$ was selected based on MAE statistics criteria (Table 3). In comparison MSE or MAPE increase the squared penalty on large errors, while MAE does not, in such model weights more errors in low concentration areas, which is key in our study. In contrast Extra Trees Regressor or Extreme Gradient Boosting models reached 0 MSE error, and over-fitted on training set, while $M_{R,MHA}$ demonstrated good generalization skill.

C. THE CHARACTERISTICS OF ALGAL BLOOMS

Most of the lakes among the analysed 105 experienced light to strong algal blooms during the years 2017-2021. In 90 lakes higher concentrations than 10 mg/m^3 was

TABLE 4. Mean values ± standard deviation of the false negatives (FN, false non-bloom), the false positives (FP, false bloom), the true negatives (TN, true non-bloom), and the true positives (TP, true bloom) in the case of the best performing classification model (Gated Residual Network with Variable Selection and Multi-Head attention layers model).

Feature	Mean FN ± sd	Mean FP ± sd	Mean TN ± sd	Mean TP ± sd
R490 (B2)	0.025 ± 0.011	0.021 ± 0.008	0.024 ± 0.012	0.027 ± 0.012
R560 (B3)	0.031 ± 0.014	0.021 ± 0.008	0.025 ± 0.013	0.035 ± 0.018
R665 (B4)	0.022 ± 0.009	0.013 ± 0.005	0.015 ± 0.007	0.026 ± 0.013
R705 (B5)	0.024 ± 0.011	0.012 ± 0.006	0.014 ± 0.007	0.031 ± 0.018
R740 (B6)	0.016 ± 0.009	0.010 ± 0.007	0.013 ± 0.008	0.018 ± 0.012
R783 (B7)	0.017 ± 0.009	0.010 ± 0.007	0.014 ± 0.009	0.019 ± 0.012
R842 (B8)	0.015 ± 0.008	0.009 ± 0.007	0.012 ± 0.009	0.017 ± 0.010
R865 (B8A)	0.014 ± 0.007	0.008 ± 0.006	0.012 ± 0.008	0.014 ± 0.009
R705/R665	1.07 ± 0.17	0.87 ± 0.18	0.94 ± 0.14	1.17 ± 0.25
R560/R490	1.24 ± 0.32	1.03 ± 0.26	1.06 ± 0.23	1.33 ± 0.36
R560/R665	1.40 ± 0.31	1.72 ± 0.52	1.68 ± 0.38	1.37 ± 0.27
R560/R705	1.32 ± 0.30	2.30 ± 1.89	1.82 ± 0.54	1.21 ± 0.31
BD	0.005 ± 0.005	0.001 ± 0.001	0.000 ± 0.001	0.009 ± 0.008
Eq_diff	18.5 ± 9.3	13.6 ± 2.8	13.6 ± 4.5	24.7 ± 16.6
AVW	559 ± 6	550 ± 7	551 ± 7	561 ± 6

TABLE 5. The regression problem models' performance metrics on the test set and a number of model parameters.

Model	H	MAE	MSE	R^2	MAPE	MSE train	Number of params
Residual DNN + MHA	128	**7.97**	**183.41**	**0.55**	2.48	259.24	1071033
Residual DNN	128	8.30	202.35	0.50	2.32	258.30	138809
Residual DNN + 2xMHA	8	8.39	187.90	0.54	2.67	286.10	10459
Gated Residual1 + VS + FC	4	8.44	188.93	0.54	**2.13**	220.44	1439
Cross Deep DNN	128	8.45	186.61	0.54	2.37	277.47	76523
Wide Deep DNN	64	8.48	192.06	0.53	2.13	256.43	21703
Gated Residual3 + VS + FC	4	8.49	194.02	0.52	2.17	230.24	1439
Gated Residual2 + VS	16	8.64	202.08	0.50	2.06	233.86	14975
Gated Residual5 + MHA	16	8.80	231.12	0.43	1.94	228.06	19487
CatBoost Regressor	-	8.83	223.47	0.45	NaN	12.22	-
Extra Trees Regressor	-	8.85	218.51	0.46	NaN	0.00	-
Random Forest Regressor	-	8.88	216.27	0.47	NaN	33.17	-
Linear Regression	-	8.89	194.85	0.52	NaN	237.54	-
Wide Deep DNN + MHA	8	8.94	216.19	0.47	1.94	248.97	49377
Gradient Boosting Regressor	-	9.01	223.03	0.45	NaN	57.47	-
Cross Deep DNN + MHA	16	9.16	224.74	0.45	1.88	258.93	194177
Ridge Regression	-	9.43	207.25	0.49	NaN	248.25	-
Light Gradient Boosting Machine	-	9.48	244.13	0.40	NaN	48.10	-
Bayesian Ridge	-	9.52	210.00	0.48	NaN	250.55	-
Elastic Net	-	9.52	210.15	0.48	NaN	250.62	-
Lasso Regression	-	9.53	210.25	0.48	NaN	250.62	-
Gated Residual4 + VS + FC	4	9.58	237.63	0.42	2.14	265.04	1643
Huber Regressor	-	9.59	237.53	0.42	NaN	259.48	-
Baseline	8	9.61	229.87	0.43	1.81	278.19	913
K Neighbors Regressor	-	9.69	245.37	0.40	NaN	202.93	-
DNN + MHA	2	9.71	225.86	0.44	2.08	254.81	9153
Extreme Gradient Boosting	-	9.76	259.66	0.36	NaN	0.09	-
Orthogonal Matching Pursuit	-	9.83	220.25	0.46	NaN	258.70	-
CNN	128	10.64	312.31	0.23	1.39	316.10	79809
Dummy Regressor	-	12.55	411.40	-0.01	NaN	414.94	-
Least Angle Regression	-	12.55	411.40	-0.01	NaN	414.94	-
Decision Tree Regressor	-	12.67	455.19	-0.12	NaN	0.00	-
AdaBoost Regressor	-	13.89	297.12	0.27	NaN	274.00	-
Passive Aggressive Regressor	-	14.87	406.54	0.00	NaN	391.44	-

observed at least once during this time period as measured in situ. Similar information was obtained using satellite and modeled data, also the time step of satellite data was more frequent as there were five to 40 observations per year in a lake. Algal blooms were identified in 92 lakes according to satellite data. The start, end dates, and duration of blooms were calculated based on the approximated Bèzier curve applied on satellite observations (Figure 4).

Shallow lakes that constituted the largest group of lakes in the dataset (59%) often experienced algal blooms. In all the lakes algal blooms were observed either using in situ (57 lakes) or satellite data (61 out of 62 lakes). The concentrations reaching as high as 175.5 mg/m^3 were observed using modeled data that was very similar to the 172.4 mg/m^3 value measured in situ. The average start of blooms was the 27th of April (date of year, doy = 117) (Table 7), nonetheless, in different years the mean start was earlier (doy = 106) – the

TABLE 6. Statistics of model error (mg/m^3) by lake type.

Class	Type	Error count	Mean	Sd	Min	25%	50%	75%	Max	MAPE, %
Non-bloom	Shallow	84	9.1	12.2	0.2	2.5	5.4	10.4	68.8	344
	Medium deep	121	5.3	4.2	0.1	2.4	4.2	7.3	23.3	221
	Deep	13	2.6	1.3	0.9	1.8	2.1	3.1	5.4	83
Bloom	Shallow	118	11.0	11.0	0.1	3.4	7.5	14.2	56.9	38
	Medium deep	96	10.2	13.0	0.0	3.5	6.9	11.8	77.3	37
	Deep	2	9.0	0.7	8.5	8.7	9.0	9.2	9.5	83

TABLE 7. Main algal bloom characteristics by lake type: N – number of year+lake combinations, mean start date (date of year, doy) of the bloom, standard deviation (Sd) of start date (days), mean end date (doy), standard deviation of bloom end date (days), mean bloom duration (days), standard deviation of bloom duration (days), mean date of chlorophyll α maximum (doy), standard deviation of chlorophyll α maximum date (days), mean maximum chlorophyll α concentration, mg/m^3, standard deviation of mean of maximum chlorophyll α concentration, mg/m^3.

Type	N	Mean start date, doy	Sd of start date, days	Mean end date doy	Sd of end date, days	Mean duration, days	Sd of duration, days	Mean date of chlorophyll α maximum, doy	Sd of date of chlorophyll α maximum, days	Mean maximum chlorophyll concentration, mg/m^3	Sd of chlorophyll α maximum, mg/m^3
Shallow	223	117	35	276	32	150	56	225	52	35.3	30.2
Med. deep	99	125	51	274	36	123	64	217	64	25.8	18.7
Deep	4.0	214	63	274	13	44.0	42	224	64	12.3	1.6

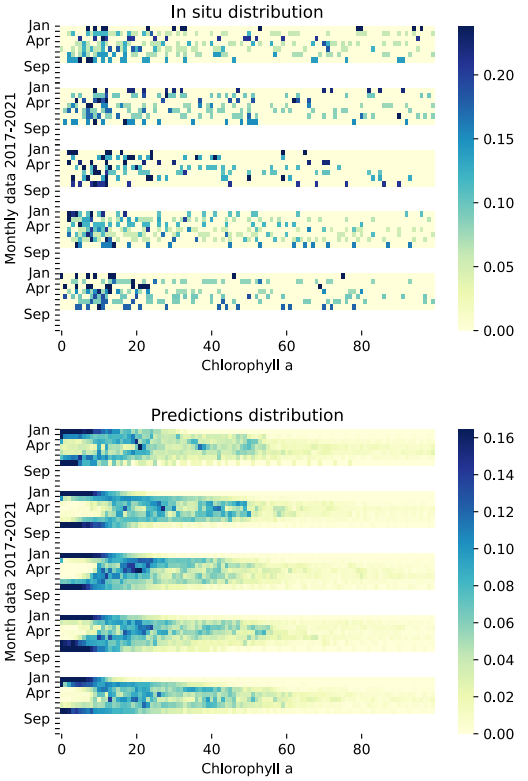

FIGURE 3. The chlorophyll α values probabilistic distribution collected in situ, and using smoothed interpolated values. The y-axis represent aggregated monthly data, x-axis chlorophyll α values. The graph panels are arranged in descending order from 2017 at the top to 2021 at the bottom.

16th of April (year 2019) or later – the 17th of May (year 2017). The mean start of the bloom in Lithuanian lakes was the 26th of April (doy = 116), while in Latvian lakes blooms started on average 2 days earlier (doy = 118), and in Estonia – 9 days later (doy = 125) than in Lithuania.

On the other hand, in shallow lakes the mean end of the bloom was the 2nd of October (doy = 276) and varied from the 14th of September (year 2020) (doy = 258) to the 14th of October (year 2019) (doy = 288). On average the end of bloom was on the 6th of October in Lithuania (doy = 280), whereas the end of bloom was earlier in Estonia (the 19th of September, doy = 262), nonetheless in Latvian lakes the blooming season ended on average a day later than in Lithuania (doy = 281). The year of 2019 was distinguished by earlier start and later end of blooms, thus the duration of blooms could have been longer by two months in lakes where the bloom was continuous. However, in some cases the blooms started later in the season, therefore, the mean bloom duration in a lake varied from 6 days to 196 days (mean = 150 days, sd = 56 days) (Figure 5).

There were 35 lakes in medium depth lake group, in 31 of them algal blooms occurred during 2017-2021 time period, while 10-26 of them were blooming during various years of 2017-2021. The start of the algal blooms in these lakes were on average seven days later than in shallow lakes (the 4th of May, doy = 125) see Table 7. Similarly to shallow lakes, in medium deep lakes earlier start of algal blooms was observed in year 2019 and later than the average in year 2018 and 2020. In Lithuanian lakes the start of algal blooms was later than on average (the 26th of May, doy = 146), while in Estonia and Latvian lakes the model showed earlier start of the blooms in mid and late April. The mean date of the end of algal blooms were earlier by two days in medium deep and deep lakes than in shallow lakes. The duration of blooms on average was shorter in medium deep lakes than in shallow lakes by 27 days (mean = 123, sd = 64) (Figure 5).

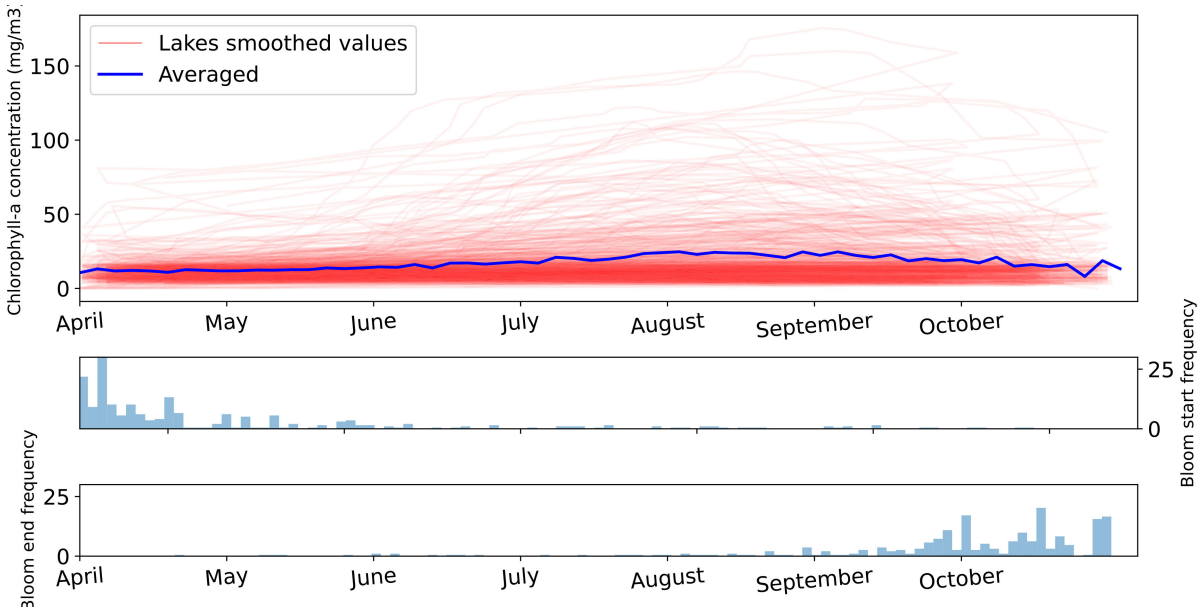

FIGURE 4. The distribution of chlorophyll α concentration, algal bloom start and end dates as determined from smoothed predictions.

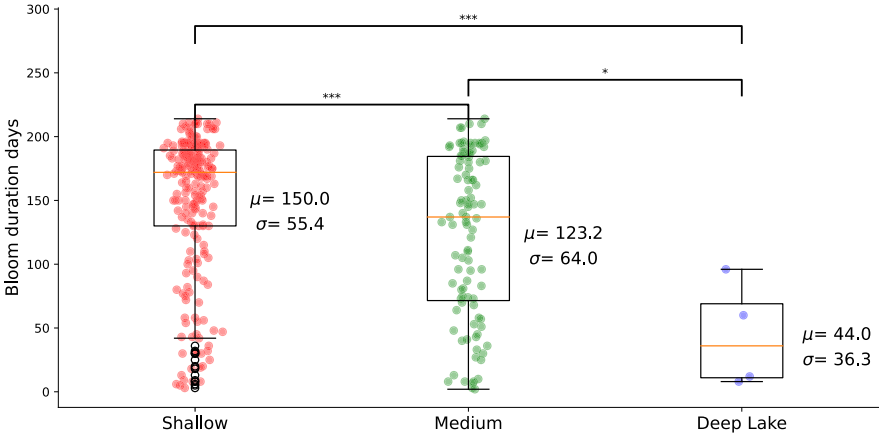

FIGURE 5. The bloomimg duration distribution based on lake type (Shallow/Medium/Deep lake). The mean and standard deviation values are presented. The differences are significant between all pairs with p-values $* - p.v < 0.05, * * * - p.v < 0.001$ using t-test.

In deep lakes the algal blooms are rather rare, five out of 8 of our analysed lakes experienced light algal blooms with concentrations reaching up to 27.2 mg/m^3. Algal blooms in these lakes occurred mostly in May and August and continued for 8 to 96 days (mean = 44, sd = 42).

The maximum chlorophyll α concentrations most often were observed in August (28.5% of bloom cases) and hereafter in September and October (18.7% of cases in each month). On average in shallow lakes the maximum concentration (mean = 35.3 mg/m^3) was reached on the 13th of August, in medium deep lakes it occurred 8 days earlier (mean = 25.8 mg/m^3 on the 5th of August), and in deep lakes a day earlier than in shallow lakes (mean = 12.3 mg/m^3 on the 12th of August).

The most intensive and the longest algal blooms were mostly observed in shallow lakes. There were 12 lakes in which during algal blooms mean chlorophyll α concentration was higher than 50 mg/m^3 and/or maximum chlorophyll α concentration was up to 100 mg/m^3 indicating intensive algal blooms. These lakes exhibited different patterns of algal bloom development. Most of these lakes experienced higher than 10 mg/m^3 (our bloom threshold) chlorophyll α concentrations already in the beginning of the season in April (Figure 6). However, in Lake Ūdrija chlorophyll α concentration was much higher already in the beginning of the warm season. The mean chlorophyll α concentration in this lake (100.4 mg/m^3) was the highest of all analysed lakes throughout the five-year period (2017-2021). Most of the lakes experienced steady increase of concentrations during April-June months and the maximum was reached in July or August.

The created models enable to obtain chlorophyll α data for all the lakes in Lithuania, Latvia, and Estonia that are routinely monitored by local environmental protection agencies. Using satellite data and models increases observational frequency from 4-6 in situ observations to up to 40 observations in a lake in a year (Figure 7).

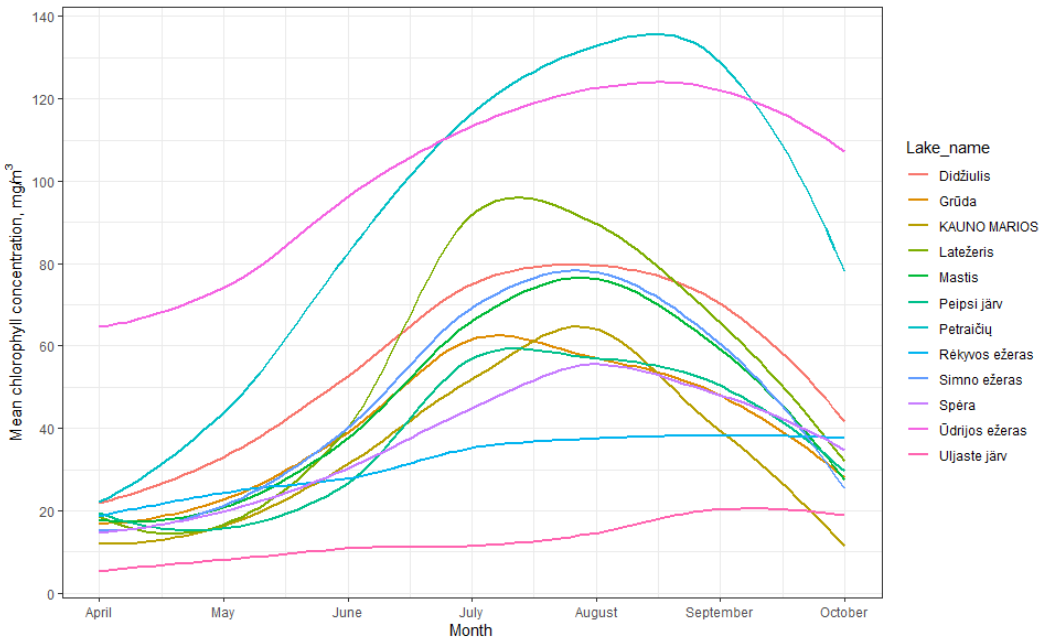

FIGURE 6. The mean chlorophyll α concentration in 12 lakes with high mean (>50 mg/m^3) and or high maximum (> 100 mg/m^3) chlorophyll α concentraion.

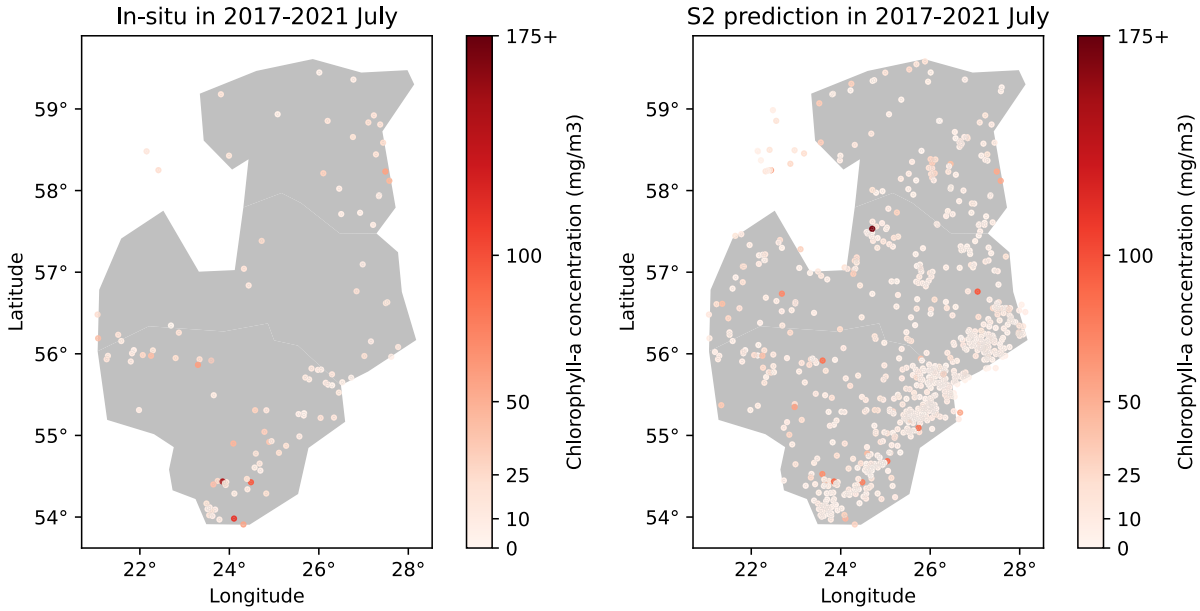

FIGURE 7. The distribution of mean chlorophyll α concentration in July (2017-2021) when algal blooms commonly occur. The concentrations obtained using in situ measurements (left panel) and retrieved from Sentinel-2 satellite and models (right panel) in Baltic states.

In addition, satellite data and created models allow us to observe spatial distribution of chlorophyll α concentration (Figure 8). Kaunas reservoir is known for strong annual cyanobacterial blooms; however, Lithuanian Environmental Protection Agency carries out measurements in one point that is in the northern part of the reservoir. Spatial information provided by satellite shows that in that area chlorophyll α concentration is often significantly lower in the monitoring site than in a few other locations in the western part and middle part of the reservoir (Figure 8). Thus, the selected monitoring site is not representative of cyanobacterial blooms in this reservoir.

IV. DISCUSSION

In this study we created a deep neural network-based classification model to separate algal blooming conditions from non-bloom conditions and a regression model to estimate the intensity of the bloom through chlorophyll α concentration. The models can be used on their own when different information may be needed – the classification when

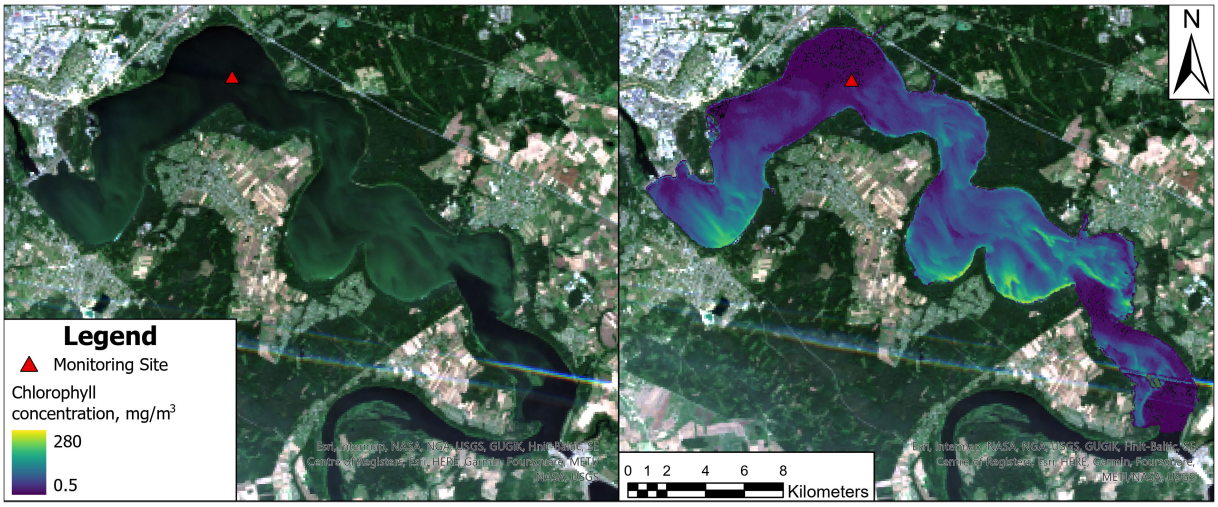

FIGURE 8. The natural colours RGB composite (left) and the spatial distribution of chlorophyll α concentration in Kaunas reservoir (area=63.5 km^2, Lithuania) on the 5th of August 2019 (right).

only the fact of bloom or non-bloom conditions is required, and regression when more specific information on algal bloom is required, such as, the intensity of bloom. In addition, the application of Bèzier curve on observations allow to gather information between the observations and obtain start and end dates, as well as, the duration of algal blooms.

Satellite data can complement the traditional in situ data with more frequent and large areas covering observations. However, the preparation of satellite retrieved dataset requires a lot of attention. Even though we prepared our dataset by removing data affected by clouds or cloud shadows and filtered it using shortwave infrared band (central wavelength 1610 nm) we came across some unusual chlorophyll α values in a few lakes in the beginning of the season. The visual check of satellite images revealed that lake Peipsi was partially frozen in April 2018 and thus our model predicted unusually high values (100 mg/m^3 and higher). This could be solved by selecting later dates; however, in this case we started in April as in the southern part of our analysed region, lakes were without ice and some of them showed as high as 50 mg/m^3 chlorophyll α concentrations. The detected unusual values were removed before further analysis. When model is applied elsewhere, it is important to examine the local conditions, such as freeze and ice break up dates. In addition, chlorophyll α model gave a few negative chla values, that were caused by very low reflectance values due to low sun angle, thus, it is important to remove these values before carrying out further analysis.

Moreover, model results might be affected by high suspended matter content in spring time, thus, in some cases the start of the bloom could have been determined earlier (for example Latvian lakes in 2019) than it started.

In this study we analysed the algal and cyanobacterial blooms without specifying whether algae or cyanobacteria are the cause of the bloom. We used Sentinel-2 MSI data that are not as good for cyanobacteria observation as other satellite sensors, such as, Ocean and Land Colour Instrument onboard Sentinel-3 that have a spectral band at the wavelengths of 620 nm for detection of absorption of cyanobacteria-specific pigment phycocyanin [46]. Moreover, since more than half of lakes in the region are smaller than 1 km^2 we used data from a satellite with higher spatial resolution (Sentinel-2) as opposed to 300 m resolution Sentinel-3. Also, in our case only concentrations of a common to both algae and cyanobacteria pigment chlorophyll α were available, thus, only for approximate estimation equations derived by other authors [52] could be used.

In this study the best classification model gave F1 score of 0.72 and regression model is able to predict chlorophyll α concentration with mean absolute error of 8.22. In other studies, such as recent Mozo et al. [39] study, authors predicting chlorophyll α concentration from water temperature, pH, electrical conductivity, and system battery obtained lower mean absolute error (3.962-5.246), and a similar F1 score of 0.737-0.667 for bloom classification (triggering the level 1 alarm). In contrast, we achieved competitive results on regression task on predicting chlorophyll α, and better blooming classification results from satellite data only. This is a significant improvement in comparison to a necessity to have in situ data like pH. In addition, we demonstrate the model performance on a much larger dataset that contains 105 lakes in comparison to one reservoir in Mozo et al. [39] study.

Our model was constructed using a relatively small dataset comprising 909 match-up points. Within the dataset, 40% exhibited concentrations surpassing 10 mg/m^3, while only 6% exceeded 50 mg/m^3 concentrations. Given the underrepresentation of high concentrations in our dataset, the addition of new observations is anticipated to enhance the overall performance of the model. Moreover, we employed an automated cloud removal methodology grounded in Sentinel-2 Level 2 scene classification. It is noteworthy that in certain

instances, this approach may not have effectively eliminated clouds or cloud shadows. Consequently, the adoption of an enhanced cloud removal algorithm holds the potential to contribute to the generation of a more representative training dataset. Moreover, we relied only on Sentinel-2 data, that has a limited spectral resolution. For future enhancements of the model, data from multiple satellites could be used; however, such approach necessitates more data preprocessing.

The creation of custom artificial neural networks for solving classification and regression tasks is a very active field of research. The creation of deep neural networks in our study was done by analysing state-of-the-art transformations in the field [23], [42]. In such we do not expand on the details of creation of the model, but rather focus on the results and the analysis of the results. The biggest challenge in the creation of the model was to find the right balance between the number of parameters in the model and the ability of generalization of the model. In our study we demonstrated that classical machine learning models like random forest or extreme gradient boosting have tendencies to overfit, and thus, we do not recommended them as a good choice for classification tasks. We trained the DNN models, using dropout techniques [9], and using small batch sizes [35] to not overfit the training set. The work can be expanded in future by incorporating additional environmental factors or adding additional spectral bands to the model from additional satellites.

The created models can be applied in other mid-latitude regions in waters with similar properties and intensities of algal and cyanobacterial blooms. In the cases of absorbing lakes, such as those with high coloured dissolved organic content, the model might underestimate the chlorophyll α and a separate model is needed for this type of lakes. The models help to collect more information on bloom start, end dates, duration, and intensity of the bloom via chlorophyll α concentration. The obtained spatial information of algal bloom extent and patterns complement in situ measurements and can help to choose the monitoring sites for in situ measurements to get a more representative data of larger water bodies.

V. CONCLUSION

Our results showed the advantage of deep neural network-based models over simpler machine learning models such as decicion trees-based models that over-fitted on our training dataset. The best performing $M_{GRN,VS,MHA}$ model with AUC $= 0.805$ was able to classify most of the algal bloom events from our dataset (84%). However, the accuracy of classifying non-bloom conditions was lower (63%) partially due to cases where waters were more complex and other optically active subtances, such as suspended matter, were present.

Our created model for estimation of intensity of algal bloom was able to retrieve chlorophyll α concentration with mean absolute error of 7.97 mg/m^3. The results are comparable to predictions of the soft sensor models, which require on site hardware and human resources.

Finally, we proposed methodology to identify blooming start and end dates by smoothing point-wise chlorophyll α predictions using Bèzier curves. While identification of algal bloom start and end dates is difficult from irregular in situ measurements, the analysis of smoothed chlorophyll α concentration predictions allow the identification including multiple algal bloom periods in a year. This led to first analysis of the blooming time comparing over Baltic region lakes. Our results also showed the impact of lake type on the blooming time.

Our results also demonstrated that experts in environmental agencies can use the proposed model on pixel-by-pixel basis (see Figure 8) to identify the blooming waters in specific lake. This could be used to identify the new pollution sources, sudden changes or for selection of monitoring sites for in situ measurements.

DATA AVAILABILITY

The data of this work were collected from Enviromental protection agencies of Baltic countries and satellite images from Sentinel-2, which are available for free. The code and data used to develop the experiments in this work is found in the following repository: https://github.com/dfrgttn/s2-for-algal-bloom.

ACKNOWLEDGMENT

The authors would like to acknowledge the Lithuanian Environmental Protection Agency, the Latvian Environment, Geology and Meteorology Centre, and the Estonian Environment Agency for providing the in situ monitoring data. The authors would like to thank to Vilnius University Science Promotion Fund for supporting initial stage of research (MSF-JM7/2021) which extended to LMTLT project and this publication. They also would like to thank Google LLC for supplying computational infrastructure for this study via the Google Earth Engine platform.

AUTHOR CONTRIBUTIONS STATEMENT

Dalia Grendaitė: Methodology on algal blooms, validation, Linas Petkevičius: Methodology on machine learning models, experiments with machine learning models, project administration, Dalia Grendaitė and Linas Petkevičius: Results analysis, visualization, writing-original draft preparation, writing-review and editing, funding acquisition. All authors have read and agreed to the published version of the manuscript.

APPENDIX
MATHEMATICAL MODELS
A. DEEP NEURAL NETWORKS

The deep neural network (DNN) is a fully-connected neural network [9], with each deep layer transformation is defined:

$$h_{l+1} = f(\theta_l h_l + b_l), \tag{1}$$

where $h_l \in \mathbb{R}^{d_l}$, $h_{l+1} \in \mathbb{R}^{d_{l+1}}$ are the l-th and $(l+1)$-th hidden layer, respectively; $\theta_l \in \mathbb{R}^{d_{l+1} \times d_l}$, $b_l \in \mathbb{R}^{d_{l+1}}$ are unknown

model parameters for the l-th deep layer, which are estimated from the data; and $f(\cdot)$ is the activation function.

B. CONVOLUTIONAL NEURAL NETWORKS

A CNN is constructed by several convolution and pooling operations, usually followed by one or more fully connected layers. The weight is a four dimensional tensor, one dimension (F) being the number of feature maps of the convolutional input layer, another (F_p) is the number of feature maps of the convolutional output layer. For a given layer n, the convolution operation:

$$a_{f\,l}^{(t)(l)} = \sum\nolimits_{f'=0}^{F_l-1} \boldsymbol{K}_{f'}^{(o)f} h_{f'}^{(t)(l)} \ ,$$

where o characterizes the $o+1$ convolution and $\boldsymbol{K}$ - learnable kernel in the CNN. Here l denotes the l'th hidden layer of the network (and thus belongs to $[\![0, N-1]\!]$), and $f \in [\![0, F_{l+1}-1]\!], l \in [\![0, N_{l+1}-1]\!]$. This followed by the activation function

$$h_f^{(t)(l+1)} = g\left(a_f^{(t)(l)}\right) \ ,$$

Finally, the output is computed as in a DNN

$$a_f^{(t)(N-1)} = \sum_{f'=0}^{F_N-1} \boldsymbol{K}_{f'}^{(o)f} h_{f'}^{(t)(N-1)} \ , h_f^{(t)(N)} = o\left(a_f^{(t)(N-1)}\right),$$

where as in a DNN, o is final activation.

C. NEURAL NETWORKS WIDENING LAYER

The wide transformation generalize the linear model of the form $y = \theta^T x + b$ [60]. Where y is the prediction, $x = [x_1, x_2, \ldots, x_d]$ is a vector of d features, $\theta = [\theta_1, \theta_2, \ldots, \theta_d]$ are the model parameters and b is the bias. The input covariates set is enriched by transformed covariates. Most often its done by cross-product transformation:

$$\phi_{new}(x) = \prod_{i=1}^{d} x_i^{c_i} \quad c_i \in \{0, 1\} \tag{2}$$

where c_i is a indicator that is 1 if the i-th feature is part of the new covariate ϕ_{new}, and 0 otherwise. In linear model it's known as covariates interactions.

1) NEURAL NETWORKS COMBINATION LAYER

The combination layer concatenates the outputs from multiple hidden layers and feed the concatenated vector to a new transformation:

$$h_{l+1} = f(\theta_l[h_{l_1}^T, h_{l_2}^T] + b_l), \tag{3}$$

where $h_{l_1} \in \mathbb{R}^{d_1}, h_{l_2} \in \mathbb{R}^{d_2}$ are the outputs from any two different hidden layers, $\theta \in \mathbb{R}^{(d_1+d_2)}$ is the new weight vector for the combination layer.

2) NEURAL NETWORKS CROSSING LAYER

The crossing transformation in neural network [60] is by applying explicit feature crossing. The crossing is composed of cross layers, with each transformations:

$$x_{l+1} = x_0 x_l^T \theta_l + b_l + x_l = f(x_l, \theta_l, b_l) + x_l, \tag{4}$$

where $x_l, x_{l+1} \in \mathbb{R}^d$ are feature vectors denoting the outputs from the l-th and $(l+1)$-th cross layers, $\theta_l, b_l \in \mathbb{R}^d$ are the unknown weights and bias parameters of the l-th layer. The cross layer additionally add its input after a covariates crossing f, where activation function $f : \mathbb{R}^d \mapsto \mathbb{R}^d$ used to the residual of $x_{l+1} - x_l$ estimation also known as residual learning [9].

3) NEURAL NETWORKS GATING LAYER

In order to determine the best individual input covariate transformation for machine learning model, some non-linear transformations can be done before passing to fully-connected layers. One of the possible input data transformation layers is Gated Residual Network (GRN) [31]. The GRN takes in a primary input x and an additional vector c:

$$\begin{aligned} \text{GRN}\,(x, c) &= \text{Norm}\,(x + \text{GLU}(h_1))\,, \\ h_1 &= \theta_1\, h_2 + b_{1,}\,, \\ h_2 &= \text{ELU}\left(\theta_2\, x + \theta_3\, c + b_{2,}\right), \end{aligned} \tag{5}$$

where ELU is the Exponential Linear Unit activation function [3], $h_1 \in \mathbb{R}^d, h_2 \in \mathbb{R}^d$ are intermediate transformations, Norm is layer normalization of [26]. The GLU:

$$\text{GLU}(\gamma) = \sigma(\theta_4\, \gamma + b_4) \odot (\theta_5\, \gamma + b_5), \tag{6}$$

where $\sigma(.)$ is the sigmoid activation function, $\gamma \in \mathbb{R}^d, \theta_{(.)} \in \mathbb{R}^{d\times d}, b_{(.)} \in \mathbb{R}^d$ are the model parameters and biases, $\odot$ is the element-wise Hadamard product, and d is the hidden state size. GLU allows to identify the significance of feature by opening/closing gate of using value based on input values, as commonly done in recurrent neural networks [9].

4) NEURAL NETWORKS VARIABLE SELECTION LAYER

Variable selection layer proposed by Lim et al. [31] measures the significance of variables by applying Softmax normalisation within the hidden layers.

Let $h^{(j)} \in \mathbb{R}^d$, be some transformation of hidden layer (or GRN), then combination of all variables followed by a Softmax layer:

$$\alpha = \text{Softmax}\left([h^{(1)}, .., h^{(j)}, .., h^{(d)}]\right), \tag{7}$$

At each variable additional layer of non-linear processing is applied by processing each $h^{(j)}$ through its own GRN:

$$\tilde{h^{(j)}} = \text{GRN}\left(h^{(j)}, c\right), \tag{8}$$

where $\tilde{h^{(j)}}$ is the processed hidden vector and c is join shared weight vector. We note that each hidden layer has its own

$\text{GRN}_{h^{(j)}}$. Processed hidden layers are then weighted by their selection weights:

$$\tilde{h} = \sum_{j=1}^{d} \alpha^{(j)} h^{\tilde{(j)}}, \tag{9}$$

where $\alpha^{(j)}$ is the j-th element of vector α.

5) MULTI-HEAD ATTENTION LAYER

The multi-head attention layers were proposed in transformer-based architectures [28], [58]. In general, attention mechanisms scale learnable values $V \in \mathbb{R}^{n \times d_V}$ based on learned dictionary relationships between keys $K \in \mathbb{R}^{n \times d}$ and input representations queries $Q \in \mathbb{R}^{n \times d}$:

$$\text{Attention}(Q, K, V) = A(Q, K)V, \tag{10}$$

where $A()$ is a scaled dot-product attention [58], Q, K, V - are linear transformation of data:

$$A(Q, K) = \text{Softmax}(QK^T/\sqrt{d}). \tag{11}$$

Equivalently to layers stacking [9], in attention the multi-head attention is proposed in [58], it repeats attention of mechanism on transform attention heads:

$$\text{MultiHead}(Q, K, V) = [H_1, \ldots, H_{m_H}]\,\theta_H, \tag{12}$$

$$H_h = \text{Attention}(Q\,\theta_Q^{(h)}, K\,\theta_K^{(h)}, V\,\theta_V^{(h)}), \tag{13}$$

where $\theta_K^{(h)} \in \mathbb{R}^{d \times d}$, $\theta_Q^{(h)} \in \mathbb{R}^{d \times d}$, $\theta_V^{(h)} \in \mathbb{R}^{d \times d_V}$ are head-specific weights for keys, queries and values, and $\theta_H \in \mathbb{R}^{(m_H \cdot d_V) \times d}$ linearly combines outputs concatenated from all heads H_h.

REFERENCES

[1] C. E. Binding, T. A. Greenberg, G. McCullough, S. B. Watson, and E. Page, "An analysis of satellite-derived chlorophyll and algal Bloom indices on lake Winnipeg," *J. Great Lakes Res.*, vol. 44, no. 3, pp. 436–446, Jun. 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0380133018300443

[2] C. Brockmann, R. Doerffer, M. Peters, S. Kerstin, S. Embacher, and A. Ruescas, "Evolution of the C2RCC neural network for Sentinel 2 and 3 for the retrieval of ocean colour products in normal and extreme optically complex waters," vol. 740, 2016, p. 54.

[3] D.-A. Clevert, T. Unterthiner, and S. Hochreiter, "Fast and accurate deep network learning by exponential linear units (ELUs)," in *Proc. ICLR*, 2016, pp. 1–16.

[4] M. M. Coffer, B. A. Schaeffer, W. B. Salls, E. Urquhart, K. A. Loftin, R. P. Stumpf, P. J. Werdell, and J. A. Darling, "Satellite remote sensing to assess cyanobacterial Bloom frequency across the United States at multiple spatial scales," *Ecol. Indicators*, vol. 128, Sep. 2021, Art. no. 107822.

[5] R. Doerffer and H. Schiller, "The MERIS case 2 water algorithm," *Int. J. Remote Sens.*, vol. 28, nos. 3–4, pp. 517–535, Feb. 2007.

[6] D. Donis et al., "Stratification strength and light climate explain variation in chlorophyll a at the continental scale in a European multilake survey in a heatwave summer," *Limnol. Oceanogr.*, vol. 66, no. 12, pp. 4314–4333, 2021.

[7] The Ministry of Environment. (2018). *Pavirsiniu Vandens Telkiniu Tipu Aprasas*. [Online]. Available: https://e-seimas.lrs.lt/portal/legalAct/lt/TAD/TAIS.256896/asr

[8] C. Fang et al., "Global divergent trends of algal blooms detected by satellite during 1982–2018," *Global Change Biol.*, vol. 28, no. 7, pp. 2327–2340, Apr. 2022.

[9] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016.

[10] N. Gorelick, M. Hancher, M. Dixon, S. Ilyushchenko, D. Thau, and R. Moore, "Google Earth Engine: Planetary-scale geospatial analysis for everyone," *Remote Sens. Environ.*, vol. 202, pp. 18–27, Dec. 2017, doi: 10.1016/j.rse.2017.06.031.

[11] D. Grendaite, "Chlorophyll-a concentration assessment in Lithuanian lakes using remote sensing," M.S. thesis, Dept. Hydrol. Climatol., Vilnius Univ., Vilnius, U.K., 2018.

[12] D. Grendaitė and E. Stonevičius, "Machine learning algorithms for biophysical classification of Lithuanian lakes based on remote sensing data," *Water*, vol. 14, no. 11, p. 1732, May 2022.

[13] B. Grizzetti, A. Pistocchi, C. Liquete, A. Udias, F. Bouraoui, and W. van de Bund, "Human pressures and ecological status of European rivers," *Sci. Rep.*, vol. 7, no. 1, pp. 1–11, Mar. 2017.

[14] M. J. Harke, M. M. Steffen, C. J. Gobler, T. G. Otten, S. W. Wilhelm, S. A. Wood, and H. W. Paerl, "A review of the global ecology, genomics, and biogeography of the toxic cyanobacterium, *microcystis* spp." *Harmful Algae*, vol. 54, pp. 4–20, Apr. 2016.

[15] D. Hermes, "Helper for bézier curves, triangles, and higher order objects," *J. Open Source Softw.*, vol. 2, no. 16, p. 267, Aug. 2017, doi: 10.21105/joss.00267.

[16] P. J. Huber, "Robust estimation of a location parameter: Annals mathematics statistics," *IEEE Trans. Signal Process.*, vol. 56, no. 6, pp. 2346–2356, Jun. 1964.

[17] J. Huisman, G. A. Codd, H. W. Paerl, B. W. Ibelings, J. M. Verspagen, and P. M. Visser, "Cyanobacterial blooms," *Nature Rev. Microbiol.*, vol. 16, no. 8, pp. 471–483, 2018.

[18] I. Ioannou, A. Gilerson, B. Gross, F. Moshary, and S. Ahmed, "Deriving ocean color products using neural networks," *Remote Sens. Environ.*, vol. 134, pp. 78–91, Jul. 2013.

[19] J. Jaagus, A. Briede, E. Rimkus, and K. Remm, "Variability and trends in daily minimum and maximum temperatures and in the diurnal temperature range in Lithuania, Latvia and Estonia in 1951–2010," *Theor. Appl. Climatol.*, vol. 118, nos. 1–2, pp. 57–68, Oct. 2014.

[20] J. Karosienė, K. Savadova-Ratkus, A. Toruńska-Sitarz, J. Koreivienė, J. Kasperovičienė, I. Vitonytė, A. Błaszczyk, and H. Mazur-Marzec, "First report of saxitoxins and anatoxin—A production by cyanobacteria from Lithuanian lakes," *Eur. J. Phycol.*, vol. 55, no. 3, pp. 327–338, Jul. 2020.

[21] N. Kashulin, T. Kashulina, and A. Bekkelund, "Long-term eutrophication and dynamics of Bloom-forming microbial communities during summer HAB in large Arctic lake," *Environments*, vol. 8, no. 8, p. 82, Aug. 2021.

[22] S. Keller, P. Maier, F. Riese, S. Norra, A. Holbach, N. Börsig, A. Wilhelms, C. Moldaenke, A. Zaake, and S. Hinz, "Hyperspectral data and machine learning for estimating CDOM, chlorophyll a, diatoms, green algae and turbidity," *Int. J. Environ. Res. Public Health*, vol. 15, no. 9, p. 1881, Aug. 2018.

[23] A. Khan, A. Sohail, U. Zahoora, and A. S. Qureshi, "A survey of the recent architectures of deep convolutional neural networks," *Artif. Intell. Rev.*, vol. 53, no. 8, pp. 5455–5516, Dec. 2020.

[24] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.

[25] T. Kutser, "Passive optical remote sensing of cyanobacteria and other intense phytoplankton blooms in coastal and inland waters," *Int. J. Remote Sens.*, vol. 30, no. 17, pp. 4401–4425, Aug. 2009.

[26] J. Lei Ba, J. Ryan Kiros, and G. E. Hinton, "Layer normalization," 2016, *arXiv:1607.06450*.

[27] P. Li, P. Hua, D. Gui, J. Niu, P. Pei, J. Zhang, and P. Krebs, "A comparative analysis of artificial neural networks and wavelet hybrid approaches to long-term toxic heavy metal prediction," *Sci. Rep.*, vol. 10, no. 1, pp. 1–15, Aug. 2020.

[28] S. Li, X. Jin, Y. Xuan, X. Zhou, W. Chen, Y.-X. Wang, and X. Yan, "Enhancing the locality and breaking the memory bottleneck of transformer on time series forecasting," in *Proc. NeurIPS*, 2019, pp. 1–16.

[29] S. Li, T. Kutser, K. Song, G. Liu, and Y. Li, "Lake turbidity mapping using an OWTs-bp based framework and Sentinel-2 imagery," *Remote Sens.*, vol. 15, no. 10, p. 2489, May 2023.

[30] M. Ligi, T. Kutser, K. Kallio, J. Attila, S. Koponen, B. Paavel, T. Soomets, and A. Reinart, "Testing the performance of empirical remote sensing algorithms in the Baltic sea waters with modelled and in situ reflectance data," *Oceanologia*, vol. 59, no. 1, pp. 57–68, Jan. 2017. https://www.sciencedirect.com/science/article/pii/S0078323416300379

[31] B. Lim, S. Ö. Arik, N. Loeff, and T. Pfister, "Temporal fusion transformers for interpretable multi-horizon time series forecasting," *Int. J. Forecasting*, vol. 37, no. 4, pp. 1748–1764, Oct. 2021.
[32] S. Lin, L. N. Novitski, J. Qi, and R. J. Stevenson, "Landsat TM/ETM+ and machine-learning algorithms for limnological studies and algal Bloom management of inland lakes," *J. Appl. Remote Sens.*, vol. 12, no. 02, p. 1, Apr. 2018.
[33] J. Llodrà-Llabrés, J. Martínez-López, T. Postma, C. Pérez-Martínez, and D. Alcaraz-Segura, "Retrieving water chlorophyll-a concentration in inland waters from Sentinel-2 imagery: Review of operability, performance and ways forward," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 125, Dec. 2023, Art. no. 103605.
[34] P. M. Maier and S. Keller, "Estimating chlorophyll a concentrations of several inland waters with hyperspectral data and machine learning models," 2019, *arXiv:1904.02052*.
[35] D. Masters and C. Luschi, "Revisiting small batch training for deep neural networks," 2018, *arXiv:1804.07612*.
[36] M. W. Matthews, "A current review of empirical procedures of remote sensing in inland and near-coastal transitional waters," *Int. J. Remote Sens.*, vol. 32, no. 21, pp. 6855–6899, 2011.
[37] T. S. Moore, M. D. Dowell, S. Bradt, and A. R. Verdu, "An optical water type framework for selecting and blending retrievals from bio-optical algorithms in lakes and coastal waters," *Remote Sens. Environ.*, vol. 143, pp. 97–111, Mar. 2014.
[38] A. Morel and L. Prieur, "Analysis of variations in ocean color," *Limnol. Oceanography*, vol. 22, no. 4, pp. 709–722, 1977.
[39] A. Mozo, J. Morón-López, S. Vakaruk, Á. G. Pompa-Pernía, Á. González-Prieto, J. A. P. Aguilar, S. Gómez-Canaval, and J. M. Ortiz, "Chlorophyll soft-sensor based on machine learning models for algal Bloom predictions," *Sci. Rep.*, vol. 12, no. 1, pp. 1–23, Aug. 2022.
[40] M. R. Ndebele-Murisa, C. F. Musil, and L. Raitt, "A review of phytoplankton dynamics in tropical African lakes," *South Afr. J. Sci.*, vol. 106, no. 1–2, pp. 13–18, Mar. 2010.
[41] C. Neil, E. Spyrakos, P. D. Hunter, and A. N. Tyler, "A global approach for chlorophyll-a retrieval across optically complex inland waters based on optical water types," *Remote Sens. Environ.*, vol. 229, pp. 159–178, Aug. 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0034425719301786
[42] Z. Niu, G. Zhong, and H. Yu, "A review on the attention mechanism of deep learning," *Neurocomputing*, vol. 452, pp. 48–62, Sep. 2021.
[43] J. E. O'Reilly, S. Maritorena, B. G. Mitchell, D. A. Siegel, K. L. Carder, S. A. Garver, M. Kahru, and C. McClain, "Ocean color chlorophyll algorithms for SeaWiFS," *J. Geophys. Res., Oceans*, vol. 103, no. C11, pp. 24937–24953, Oct. 1998.
[44] R. L. Oliver and G. G. Ganf, "Freshwater blooms," in *The Ecology of Cyanobacteria*. Dordrecht, The Netherlands: Springer, 2002, pp. 149–194, doi: 10.1007/0-306-46855-76.
[45] E. Papathanasopoulou et al., "Satellite-assisted monitoring of water quality to support the implementation of the water framework directive," EOMORES White Paper, p. 28, 2019, doi: 10.5281/zenodo.3463050.
[46] R. Pérez-González, X. Sória-Perpinyà, J. M. Soria, J. Delegido, P. Urrego, M. D. Sendra, A. Ruíz-Verdú, E. Vicente, and J. Moreno, "Phycocyanin monitoring in some Spanish water bodies with Sentinel-2 imagery," *Water*, vol. 13, no. 20, p. 2866, Oct. 2021. [Online]. Available: https://www.mdpi.com/2073-4441/13/20/2866
[47] A. Rakko, R. Laugaste, and I. Ott, "Algal blooms in Estonian small lakes," in *Algal Toxins: Nature, Occurrence, Effect and Detection*. Dordrecht, The Netherlands: Springer, 2008, pp. 211–220.
[48] S. Ravuri, K. Lenc, M. Willson, D. Kangin, R. Lam, P. Mirowski, M. Fitzsimons, M. Athanassiadou, S. Kashem, S. Madge, R. Prudden, A. Mandhane, A. Clark, A. Brock, K. Simonyan, R. Hadsell, N. Robinson, E. Clancy, A. Arribas, and S. Mohamed, "Skilful precipitation nowcasting using deep generative models of radar," *Nature*, vol. 597, no. 7878, pp. 672–677, Sep. 2021.
[49] T.-C. Ri and J.-S. Jo, "A genetic algorithm-optimized neural network for chlorophyll a estimation using MODIS satellite data in coastal water: Application to the sinpho bay of DPR Korea," *J. Indian Soc. Remote Sens.*, vol. 51, no. 7, pp. 1541–1551, Jul. 2023.
[50] V. Sagan, K. T. Peterson, M. Maimaitijiang, P. Sidike, J. Sloan, B. A. Greeling, S. Maalouf, and C. Adams, "Monitoring inland water quality using remote sensing: Potential and limitations of spectral indices, bio-optical simulations, machine learning, and cloud computing," *Earth-Sci. Rev.*, vol. 205, Jun. 2020, Art. no. 103187.
[51] T. W. Sederberg and T. Nishita, "Curve intersection using Bézier clipping," *Comput.-Aided Design*, vol. 22, no. 9, pp. 538–549, Nov. 1990.
[52] X. Sòria-Perpinyà, E. Vicente, P. Urrego, M. Pereira-Sandoval, C. Tenjo, A. Ruíz-Verdú, J. Delegido, J. M. Soria, R. Peña, and J. Moreno, "Validation of water quality monitoring algorithms for Sentinel-2 and Sentinel-3 in Mediterranean inland waters with in situ reflectance data," *Water*, vol. 13, no. 5, p. 686, Mar. 2021. [Online]. Available: https://www.mdpi.com/2073-4441/13/5/686
[53] E. Spyrakos et al., "Optical types of inland and coastal waters," *Limnol. Oceanogr.*, vol. 63, no. 2, pp. 846–870, 2018.
[54] A. Tanaka, M. Kishino, R. Doerffer, H. Schiller, T. Oishi, and T. Kubota, "Development of a neural network algorithm for retrieving concentrations of chlorophyll, suspended matter and yellow substance from radiance data of the ocean color and temperature scanner," *J. Oceanogr.*, vol. 60, no. 3, pp. 519–530, 2004.
[55] K. Toming, T. Kutser, A. Laas, M. Sepp, B. Paavel, and T. Nõges, "First experiences in mapping lake water quality parameters with Sentinel-2 MSI imagery," *Remote Sens.*, vol. 8, no. 8, p. 640, Aug. 2016.
[56] D. Vaičiūtė, M. Buč as, M. Bresciani, T. Dabulevičienė, J. Gintauskas, J. Mėžinė, E. Tiškus, G. Umgiesser, J. Morkūnas, F. De Santi, and M. Bartoli, "Hot moments and hotspots of cyanobacteria hyperblooms in the curonian lagoon (SE Baltic Sea) revealed via remote sensing-based retrospective analysis," *Sci. Total Environ.*, vol. 769, May 2021, Art. no. 145053.
[57] R. A. Vandermeulen, A. Mannino, S. E. Craig, and P. J. Werdell, "150 shades of green: Using the full spectrum of remote sensing reflectance to elucidate color shifts in the ocean," *Remote Sens. Environ.*, vol. 247, Sep. 2020, Art. no. 111900.
[58] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. NIPS*, 2017, pp. 1–55.
[59] K. P. Wai, M. Y. Chia, C. H. Koo, Y. F. Huang, and W. C. Chong, "Applications of deep learning in water quality management: A state-of-the-art review," *J. Hydrol.*, vol. 613, Oct. 2022, Art. no. 128332.
[60] R. Wang, B. Fu, G. Fu, and M. Wang, "Deep cross network for ad click predictions," in *Proc. ADKDD*. New York, NY, USA: ACM, 2017, pp. 1–7.

DALIA GRENDAITĖ received the Ph.D. degree in physical geography from the Institute of Geosciences, Vilnius University, in 2023.

Her work encompasses remote sensing and hydrology. Using advanced remote sensing techniques, she is dedicated to understanding and monitoring the ecological dynamics of these freshwater ecosystems. Her expertise includes the assessment of water quality, ecological monitoring, and the analysis of aquatic pollution.

LINAS PETKEVIČIUS (Member, IEEE) received the Ph.D. degree in informatics from the Institute of Computer Science, Vilnius University, in 2020. Since 2022, he has been the Head of the Software Engineering Department, Institute of Computer Science. His research interests include computer vision and deep learning as well as statistical inference and outliers detection.