Multi-step Ultraviolet Index Forecasting using Long Short-Term Memory Networks

Pedro Oliveira[0000−0001−7143−5413], Bruno Fernandes[0000−0003−1561−2897], Cesar Analide[0000−0002−7796−644X] , and Paulo Novais[0000−0002−3549−0754]

Department of Informatics, ALGORITMI Centre, University of Minho, Braga, Portugal
poliveira199208@gmail.com, bruno.fmf.8@gmail.com, analide@di.uminho.pt, pjon@di.uminho.pt

Abstract. The ultraviolet index is an international standard metric for measuring the strength of the ultraviolet radiation reaching Earth's surface at a particular time, at a particular place. Major health problems may arise from an overexposure to such radiation, including skin cancer or premature ageing, just to name a few. Hence, the goal of this work is to make use of Deep Learning models to forecast the ultraviolet index at a certain area for future timesteps. With the problem framed as a time series one, candidate models are based on Recurring Neural Networks, a particular class of Artificial Neural Networks that have been shown to produce promising results when handling time series. In particular, candidate models implement Long Short-Term Memory networks, with the models' input ranging from uni to multi-variate. The used dataset was collected by the authors of this work. On the other hand, the models' output follows a recursive multi-step approach to forecast several future timesteps. The obtained results strengthen the use of Long Short-Term Memory networks to handle time series problems, with the best candidate model achieving high performance and accuracy for ultraviolet index forecasting.

Keywords: Deep Learning, Ultraviolet Index, Long Short-Term Memory Networks, Time Series Forecasting.

1 Introduction

Ultraviolet (UV) index is a standard metric used to express the magnitude of UV radiation reaching Earth's surface at a particular time, at a given region. Ozone in the stratosphere, also known as "good" ozone, protects life from harmful UV radiation. However, due to the burning of fossil fuels, which releases carbon into the atmosphere, the ozone layer has become thinner, leading to dangerous UV radiation reaching Earth's surface [1].

Over the years, the increase of UV radiation reaching Earth's surface has been associated with increased rates of skin cancers, particularly melanomas [2]. Indeed, information regarding UV index variations can be essential for the human

being. Despite harmful in high concentrations, UV radiation is also essential for humanity. As an example, exposure to UV radiation activates Vitamin D, a key regulator in the calcium and phosphate homeostasis with implications in several human body systems [2]. For a better perception of which concentrations lead to harmful UV radiation, the World Health Organisation and the World Meteorological Organisation proposed a standardised global UV Index scale. UV index values between 0 and 2 are of low risk; between 3 and 5 are of moderate risk; between 6 and 7 start carrying some risk; between 8 and 11 are very dangerous to the human being; and values higher than 11 are of extreme danger [3].

Knowing, beforehand, when the UV index will achieve high or extreme values is of the utmost importance as it allows one to adjust his behaviour and avoid risky moves. Hence, the goal of this work is to make use of Deep Learning models to forecast the UV index at a certain area for several future timesteps, in particular for the next three days. With the use of Deep Learning models, it becomes possible to forecast future time points in a given scope. Being this a time series problem, uni and multi-variate Long Short-Term Memory networks (LSTMs), a subset of Recurrent Neural Networks (RNNs), were conceived and evaluated, with the goal being to forecast the UV index.

The remainder of this paper is structured as follows: the next section summarises the state of the art on the subject of UV forecasting. The third section aims to present the materials and methods, focusing on the collected dataset, its exploration and all the applied treatments. The fourth and fifth sections yield a description of the performed experiences and a discussion of the obtained results, respectively. The sixth and final section notes major conclusions taken from this work and presents future perspectives.

2 State of the Art

Across the years several studies have been carried out on the topic of UV forecasting [4–6]. Gómez et al. [4], used the Santa Barbara DISTORT Atmospheric Radiative Transfer (SBDART) algorithm, developed by Ricchiazzi et al. [7], to predict the UV index for Valencia, Spain, in a period of 3 days. The SBDART model calculates the radiative transfer parallel to the Earth's atmosphere [7]. In this study, this model had, as input, the Total Ozone Columns (TOC) data through the Global Forecast System (GFS). Hence, as the UV incidence forecast would be for the next 3 days, there was a gap of 4 days between the last available data, also limiting a forecast to the next day. The metrics used to evaluate the models were the Root Mean Square (RMS) and Mean Value (MV). With these metrics, the authors concluded that the TOC GFS obtained interesting results when forecasting the UV index for the next day.

Another study was conducted by Ravinesh et al. [5], with the authors focusing on a short term forecast of 10 minutes. These authors propose an Extreme Machine Learning (EML) model that is based on a Single Layer Feed Forward Network (SLFN), that use a FeedForward Back Propagation (FFBP). To evaluate the performance of the models, the authors used two metrics as

criteria: Mean Absolute Error (MAE) and Root Mean Square Error (RMSE). In this study, the EML model outperformed the Multivariate Adaptive Regression Splines (MARS) model and a Hierarchical model (M5 Model Tree). The data was partitioned with 80% for training and the remaining 20% for testing and validation. The ELM model shows a decrease of 0.1, in both metrics, compared to the other models.

Puerta et al [6] proposed a Deep Learning Model to forecast UV index to predict erythema formation in the human skin. Erythema corresponds to redness of the skin due to dilation of the superficial capillaries. The fit of the model is carried out using a Deep Belief Network (DNB). In this model, the backpropagation method is also applied, adjusting the weight values according to the Mean Squared Error (MSE). The conceived neural network has, as input, the average temperature, the clear sky index, the insolation, and the UV index of the day before the one it intends to predict. Regarding the validation of the model, it was carried out by comparison with the records extracted from the National Aeronautics And Space Administration (NASA) Prediction Of Worldwide Energy Resource. A value of 66.8% of correct classifications was obtained in the values predicted by the conceived model.

Interestingly, in the environmental sustainability domain, there is a clear lack of studies that use machine or deep learning to predict the UV index. There are however studies that focuses on related topics such as forecasting ozone values at ground level. Ghoneim et al. [8], conducted a study to predict ozone concentration in a smart city. This study aimed to compare a deep learning model to regression techniques, such as Support Vector Machines (SVM). In this study, data related to pollution and weather from the CityPulse project [9] were used. Grid search was used to optimise hyperparameters of the model, such as the number of hidden layers and the number of neurons in each layer. MSE, RMSE and MAE were the used metrics. With a focus on RMSE, the deep learning model had a lower RMSE when compared with the other models. In fact, the deep learning model outperformed all other models for all used metrics.

3 Materials and Methods

The materials and methods used in this study to conceive and evaluate the candidate models are detailed in the next lines. The used evaluation metrics are also described as well as the conceived deep learning models.

3.1 Data Collection

The UV index dataset used in this study was created from scratch using real-world data. For that purpose, a software was developed to collect data from a set of soft sensors. In this case, the Open Weather Map API, whose features are the type of pollutant, the name of the city, the value of the UV index, the source from which the record is taken, and the timestamp. Data collection started on July 24, 2018 and was maintained, uninterruptedly, until the present. The developed

software makes API calls every hour using an HTTP Get request. It parses the received JSON object and saves the records on a database. The software was developed so that any other type of polluter can be easily added to the list. The present work analyses data collected until February 24, 2020.

3.2 Data Exploration

The collected dataset consists of a total of 16375 timesteps with data being physically gathered by three distinct hard sensors. Each observation, from now on designated as timestep, consists of a total of 8 features. Most features are of the string type, with the UV index being a double-typed feature. The remaining features are integers. The creation date feature consists of the date and time. Table 1 describes the features available in the collected dataset.

Table 1: Features of the collected dataset.

#	Features	Description
1	*id*	Record identifier
2	*pollution_type*	Pollutant type
3	*city_name*	Name of the city under analysis
4	*value*	UV index index
5	*data_precision*	Precision of the value feature
6	*source_name*	Source from where the timestep was obtained
7	*last_update*	Last date when the value was updated
8	*creation_date*	Timestamp (YYYY-MM-DD HH24:MI:SS)

After an analysis of the dataset, it is possible to understand that the *last update* feature does not have any value assigned as well as the *data precision* feature which is always filled with the same value. Being this a time series problem, one must be aware of all missing timesteps. In total there were missing 102 timesteps. Some of which corresponded to periods of roughly 1 month, between December 13, 2018 to January 14, 2019, and between March 7, 2019 to April 9, 2019.

To understand the variation of the UV index over the course of a year, the monthly average was analysed. Figure 1 illustrates the average of the UV index for the years 2018 and 2019. It is possible to verify that the peak of the UV index is reached during July. From that month onwards, the index declines until December and starts to increase again in January until reaching its peak. Therefore, it is possible to highlight the existence of seasonality as well as cyclicality. The highest values are reached during the summer, reducing during the fall and the beginning of the winter.

3.3 Data Preparation

The available dataset includes observations from July 24^{th}, 2018 to February 24^{th}, 2020 made at one hour time intervals. The first step in the preparation

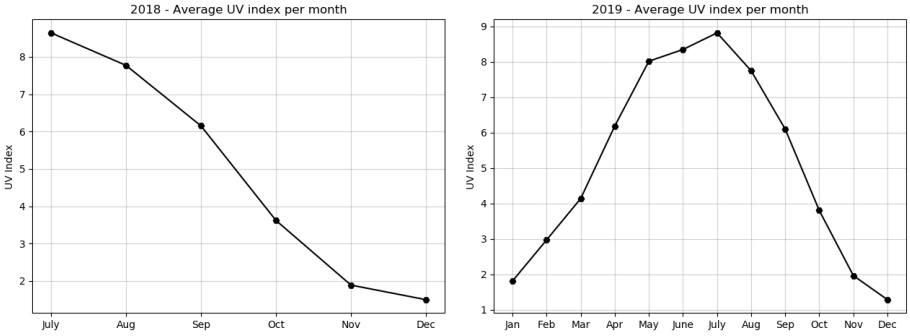

Fig. 1: Average UV index per month in the years of 2018 and 2019.

of the dataset is to apply feature engineering and create the year, month, day, hour, and minutes as features to each observation. Daily forecast requires daily UV index measures. Thus, the average of observations was considered to create daily timesteps.

No missing values were found in the dataset. However, missing timesteps were present in the dataset, either due to certain limitations of the API or because it was unavailable (not recorded or not measured). The lack of timesteps can result in the development of incorrect standards, so it was necessary to fill in the missing values. For that purpose, the Open Weather Map Historical UV API was used. By the end of this step, the dataset consisted of 581 daily timesteps.

The next step consists in removing some informative features that will not be used by the conceived models such as the *id*, *pollution type*, *city name*, *data precision*, *source name* and *last update*. The *hour* and *minute* features were also disregarded since the dataset was grouped by day.

Since LSTMs work internally with the hyperbolic tangent, normalisation of the dataset was performed, with all features falling within the range [-1,1], according to the following equation:

$$\frac{x_i - min(x)}{max(x) - min(x)} \tag{1}$$

At the end of the data preparation process, two datasets were created. Both datasets contain 581 timesteps, varying only in the number of features. One dataset (**Uni-variate**) contains only the value of the UV index for each timestep. The second dataset (**Multi-variate**) contains, beside the UV index for each timestep, the month of the year and the day.

3.4 Evaluation Metrics

To obtain the best combination of parameters of the candidate models, two error metrics were used. One, the RMSE, is a measure of accuracy, as it measures

the difference between the values predicted by the model and the true values observed. RMSE equation is as follows:

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}} \tag{2}$$

The second metric, the MAE, is the mean of the differences between predicted and observed values. Its use is mainly to complement and strengthen the confidence of the obtained values. Its equation is as follows:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{3}$$

3.5 LSTMs

RNNs constitute a class of artificial neural networks where the evolution of the state depends on the current input as well as the input at previous timesteps. This property makes it possible to carry out context-dependent processing, allowing long-term dependencies to be learned [10]. A recurrent network can have connections that return from the outgoing nodes to the incoming nodes, or even arbitrary connections between any nodes.

LSTMs are a type of RNN architecture. Unlike a traditional neural network, this architecture is used to learn from experience how to classify, process and predict time series, as is the case with this study. This type of network aims to help preserve the error, which can be propagated through time and layers. The technique of keeping the error constant, allows this type of networks to continue their learning process over many "steps" in time [10].

LSTMs contain information outside the normal flow of the recurring network, more specifically in a gated cell. Information can be stored, written or read from a given cell, in an approach similar to data in a computer's memory. These networks are used to process, predict and classify based on time series data.

4 Experiments

To achieve the objective of predicting the UV index, it was necessary to develop and tune several candidate LSTM models. The conceived models forecast the UV index for three consecutive days, following a recursive multi-step approach, i.e., predicting recursively each future timestep until it achieves the time window of three predicted days. With this multi-step approach, the prediction for timestep t is used as input to predict timestep $t + 1$.

Several experiments were carried out to find the best combination of hyperparameters for both the uni-variate and the multi-variate approaches. Performance was compared in terms of error-based accuracy, for both the uni-variate and multi-variate candidate LSTM models. Regarding the first, it uses only one feature as input, unlike the second which takes into account several features. In fact, the uni-variate models use only the UV index *value* feature to recursively

predict the future values of this index. On the other hand, the multi-variate model uses the UV index *value* as well as the *month* and *day* features, giving a stronger temporal context to the network. Regarding the searched hyperparameter configuration, they were identical between both approaches, being described in Table 2 which summarises the parameter searching space considered for uni and multi-variate models.

Table 2: Uni-Variate vs Multi-Variate hyperparameters' searching space.

Parameter	**Uni-Variate**	**Multi-Variate**	**Rationale**
Epochs	[150, 300]	[300, 500]	-
Timesteps	[7, 14]	[7, 14]	Input of 1 and 2 weeks
Batch size	[16, 23]	[16, 23]	Batch of 2 to 3 weeks
LSTM layers	[3, 4]	[3, 4]	Number of LSTM layers
Dense Layers	1	1	Number of dense layers
Dense Activation	[ReLU, tanh]	[ReLU, tanh]	Activation function
Neurons	[32, 64, 128]	[32, 64, 128]	For dense and LSTM layers
Dropout	[0.0, 0.5]	[0.0, 0.5]	For dense and LSTM layers
Learning rate	callback	callback	Keras callback
Multisteps	3	3	3 days forecasts
Features	1*	3**	Used features
CV Splits	3	3	Time series cross-validator

* Used features: *UV index*
**Used features: *UV index*, *month* and *day*

Knime was the platform used for data exploration. Python, version 3.7, was the used programming language for data preparation, model development and evaluation. Pandas, NumPy, scikit-learn and matplotlib were the used libraries. Tensorflow v2.0.0 was used to implement the deep learning models. Tesla T4 GPUs were used as well as CUDNNLSTM layers for optimized performance in a GPU environment. All hardware was made available by Google's Colaboratory, a free python environment that runs entirely in the cloud.

5 Results and Discussion

Being this a time series forecasting problem, one particular time series cross validator was used, being entitled as TimeSeriesSplit. For each prediction of each split of this cross validator, RMSE and MAE were calculated to be able to evaluate the best set of parameters. The experiments carried out for both uni and multi-variate candidate models made it clear that a stronger temporal context results in an overall decrease of both error metrics even though the best uni-variate models has a lowest MAE than the best multi-variate one. Table 3 depicts the top 3 results for both approaches.

The best LSTM model concerning the uni-variate model had an RMSE of 0.325 and an MAE of 0.236. On the other hand, the RMSE was 0.306 and the

Table 3: Uni-Variate vs Multi-Variate LSTMs top-three results.

#	Timesteps	Batch	Layers	Neurons	Dropout	Act.	RMSE	MAE
Recursive Multi-Step Uni-Variate								
116	7	16	4	64	0.5	tanh	**0.325**	**0.236**
108	7	16	3	128	0.5	tanh	0.349	0.271
8	7	16	3	64	0.5	tanh	0.354	0.26
Recursive Multi-Step Multi-Variate								
31	14	16	3	64	0.0	tanh	**0.306**	**0.249**
125	14	16	3	64	0.0	relu	0.339	0.284
73	14	23	3	32	0.0	relu	0.34	0.275

MAE was 0.249 for the best multi-variate model. Since both metrics are in the same unit of measurement as the UV index, an error of 0.3 shows that it is possible to forecast, very closely, the expected UV index for the next three days. In the multi-variate model, the inclusion of the month and day yields more accurate predictions in comparison to the uni-variate one. Interestingly, the number of inputs of the model is directly proportional to the number of timesteps, i.e., more features as input lead to an increase of the number of timesteps that are required to build a sequence. This is shown by Table 3, with the best uni-variate models requiring sequences of 7 timesteps (a week), while the best multi-variate ones require sequences of 14 timesteps (two weeks). The number of epochs and batch size was 300 and 16, respectively, for both models, with a early stopping callback stopping training when the monitored loss stopped improving. Concerning the number of hidden layers, the best uni-variate model required a more complex architecture of 4 hidden layers to achieve similar performances to the multi-variate ones, who required a shallow architecture. In addition, this shallow architecture ruled out the use of dropout, which was required by the best uni-variate models. Figure 2 presents the architectures of the best multi-variate model. Moreover, Figure 3 illustrates six predictions of three days for the best Multi-Variate LSTM model, showing very close results to the known UV index value.

6 Conclusions

Over the past few years, skin cancer prevention campaigns have increased worldwide. Knowing that exposure to ultraviolet radiation is one of the main causes for such disease, forecasting the UV index assumes particular importance. Hence, this study focused on using deep learning models, in particular LSTMs, to forecast the UV index for the next three days. Multiple experiments were performed, using a wide combination of hyperparameters for all the candidate models. The model with the best accuracy in the prediction of the UV index was the Recursive Multi-Step Multi-Variate model with a RMSE of 0.306 and a MAE of 0.249, which depict that it is possible to forecast, with very accurate results, the UV index for the next few days. Nevertheless, the models that had only the

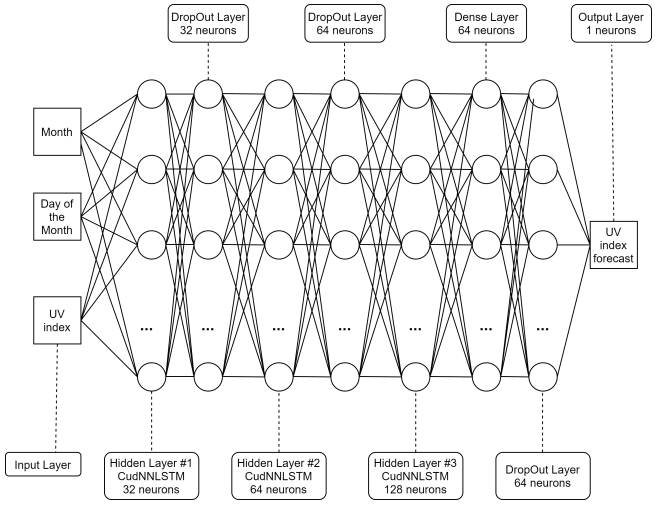

Fig. 2: Architecture of the best multi-variate model.

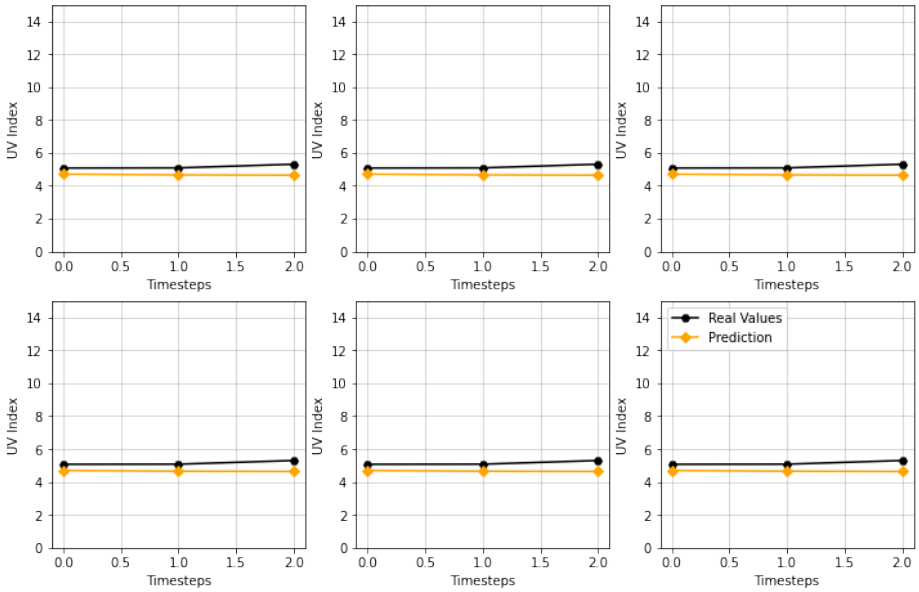

Fig. 3: Six random predictions of the best multi-variate LSTM model.

UV index as input also presented interesting results. As expected, the number of input features impacts the models' accuracy. Yet it is interesting to note that the increase in input features led to an increase of the required timesteps as well as to shallow networks.

The obtained results show promising prospects for UV index forecasting. Hence, future work will consider the inclusion of more input features such as the temperature, ozone levels and the position of the sun expressed in terms of solar zenith angle. In addition, future work will also focus on different state-of-the-art recurrent networks such as the Gated Recurrent Unit network.

Acknowledgments. This work has been supported by FCT - Fundação para a Ciência e a Tecnologia within the R&D Units project scope UIDB/00319/2020 and DSAIPA/AI/0099/2019. The work of Bruno Fernandes is also supported by a Portuguese doctoral grant, SFRH/BD/130125/2017, issued by FCT in Portugal.

References

1. Lickley, M., Solomon, S., Fletcher, S., Velders, G., Daniel, J., Rigby, M., Montzka, S., Kuijpers, L., Stone, K.: Quantifying contributions of chlorofluorocarbon banks to emissions and impacts on the ozone layer and climate. Nature Communications vol.**10**(1), 1–11 (2020)
2. Young, A., Narbutt, J., Harrison, G., Lawrence, K., Bell, M., O'Connor, C., Olsen, P., Grys, K., Baczynska, K., Rogowski-Tylman, M., and others: Optimal sunscreen use, during a sun holiday with a very high ultraviolet index, allows vitamin D synthesis without sunburn. British Journal of Dermatology vol.**181**(5), 1052–1062 (2019)
3. World Health Organization and International Commission on Non-Ionizing Radiation Protection and others: Global solar UV index: a practical guide. World Health Organization, (2002)
4. Gómez, I., Marín, M., Pastor, F., Estrela, M.: Improvement of the Valencia region ultraviolet index (UVI) forecasting system. Computer & Geosciences vol.**41**, 72–82 (2012)
5. Deo, R., Downs, N., Parisi, A., Adamowski, J., Quilty, J.: Very short-term reactive forecasting of the solar ultraviolet index using an extreme learning machine integrated with the solar zenith angle. Environmental research vol.**155**, 141–166 (2017)
6. Barrera, J., Hurtado, D., Moreno, R.: Prediction system of erythemas for phototypes i and ii, using deep-learning. Vitae vol.**22**(3), 189–196 (2015)
7. Ricchiazzi, P., Yang, S., Gautier, C., Sowle, D.: SBDART: A research and teaching software tool for plane-parallel radiative transfer in the Earth's atmosphere. Bulletin of the American Meteorological Society vol.**79**(10), 2101–2114 (1998)
8. Ghoneim, O., Manjunatha, B., and others: Forecasting of ozone concentration in smart city using deep learning. In: 2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI), pp. 1320–1326. IEEE, (2017)
9. Barnaghi, P., Tönjes, R., Höller, J., Hauswirth, M., Sheth, A., Anantharam, P.: Citypulse: Real-time iot stream processing and large-scale data analytics for smart city applications. In: Europen semantic web conference (ESWC), (2014)
10. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation vol.**9**(8), 1735–1780 (1997)