A Multi-class Classification Approach for Weather Forecasting with Machine Learning Techniques

Elias Dritsas, Maria Trigka
Department of Electrical and Computer Engineering,
University of Patras, Patras, Greece
{dritsase, trigka}@ceid.upatras.gr

Phivos Mylonas
Department of Informatics
Ionian University, Corfu, Greece
fmylonas@ionio.gr

***Abstract*—Weather forecasting is vital as extreme weather events can cause damage and even death. The science of meteorology in recent decades has made spectacular progress resulting in more reliable forecasts. Although meteorologists now have adopted modern tools for accurate weather forecasting, extreme and sudden climate changes in the atmosphere have posed accurate weather forecasting even more valuable. In this research paper, we present a multi-class classification methodology from machine learning (ML) in order to predict the five classes of weather conditions. Specifically, the One-Against-One (OAO) and One-Against-All (OAA) strategies are evaluated under Support Vector Machine (SVM) and Logistic Regression (LR) assuming, for comparison, Random Forest (RF) and k-Nearest Neighbours (k-NN). The prevailing model is linear SVM under the OAO method achieving the average Accuracy, Precision, Recall, F-Measure and Area Under Curve (AUC) of 96.64%, 96.8%, 96.6%, 96.6% and 98.5%, respectively.**

***Index Terms*—Weather Forecasting, Machine Learning, Classification, Data Analysis**

I. Introduction

Weather forecasting is the knowledge of weather change after a certain period of time (through so-called barometric systems). The process of weather forecasting is serious for inhabitants worldwide. Weather also affects businesses in various industries (such as manufacturing, retail, entertainment, agriculture and livestock) in many ways. A typical example is in agriculture, where weather affects the growth, harvesting and distribution of crops, vegetables and fruits thereby affecting the availability and variety of food. Climate, weather, and changing seasons affect people's habits by allowing for many different types of work, activities, celebrations, recreation, eating, etc. In addition to humans, plants and animals change functions according to the seasons. Finally, weather affects peoples' psychology. It has been observed that the climate and weather conditions of an area have an impact on the psychology of the inhabitants which in some cases can affect their psychology and lead to depression [1]–[4].

Meteorology is the science that studies atmospheric phenomena. Its purpose is the precise observation and monitoring of weather conditions and their description both quantitatively and qualitatively. In addition, it includes the interpretation of these phenomena and the formulation of laws governing their changes and the development of methods for accurate weather forecasting [5], [6].

Meteorologists have developed mathematical models based on the knowledge of atmospheric physics and mathematics. In this way, they aim to predict every change that takes place in the atmosphere and follows certain physical laws that can be determined by mathematical equations. Specifically, weather prediction could be made based on atmospheric data involving temperature, precipitation, humidity, wind speed, and direction [7], which may gradually change. In addition to the above, the forecast is connected with probability and mathematical models can indicate the most probable weather condition [8], [9].

Meteorological models are generally large-scale and thus time-consuming since their execution requires several procedures and calculations of a huge amount of data. The results of the meteorological models should be given in a short period of time to the departments that undertake the processing so that the final conclusions of the forecast can be made in time. Nowadays, the available computing capabilities, tools and infrastructure paved the way for the development and implementation of atmospheric models that could simulate the atmosphere and make a reliable forecast of weather [10], [11].

From a statistical signal processing perspective, machine learning techniques have played an important role in prediction tasks and thus found application in weather forecasting as well. Various computational models have been proposed by researchers that achieve a high level of accuracy [12]. The main contribution of this paper is i) a multi-class classification methodology based on the OAO and OAA following the main preprocessing steps from ML and considering SVM and LR as base models [13], [14] and ii) performance evaluation under various metrics.

The rest of the paper is organized as follows. Section II describes the relevant works with the subject under consideration. Besides, in Section III, a dataset description and analysis of the methodology followed are made. In addition, in Section IV, we discuss the acquired research results. Finally, conclusions and future directions are outlined in Section V.

II. Related work

This section will provide a brief overview of the most recent works on weather forecasting using various machine-learning

978-1-6654-8727-6/22/$31.00 © 2022 IEEE

techniques.

First, in [12], a low-cost and portable solution for weather prediction is devised. Random forest classifier provides deep insights into the data dependencies. In [15], a hybridized model was developed to predict the temperature and humidity and analyze future weather conditions. The algorithm used to train the model was weighted k-NN with k equal to 17. The accuracies were 76.30% and 46.29% during training and testing, respectively.

Moreover, in [16], a Hadoop-based weather forecasting model is proposed for efficient processing and prediction of weather data. The authors concluded that the ANFIS method predicts weather data more accurately. In [17], the authors proposed an ML forecasting model for the accurate prediction of qualitative weather information on winter precipitation types, utilized in the Apache Spark Streaming distributed framework. Three different categorization methods, such as the Bayesian, decision trees, and meta/ensemble methods were applied.

In [18], the performance of data mining and machine learning techniques using Support Vector Regression (SVR) and Artificial Neural Networks (ANN) for a robust weather prediction purpose is outlined. The data were collected from Bangladesh Meteorological Department (BMD). The SVR outperforms the ANN in rainfall prediction, and ANN achieves better results than the SVR.

The authors in [19] aim to classify the weather based on Twitter text data by using Support Vector Machine (SVM), Multinomial Naive Bayes (MNB), and Logistic Regression (LR) methods. The experimental results showed that the SVM outperforms with an accuracy value of 93% concerning the other compared methods.

In addition, a linear regression model and a variation on a functional regression model were used in [20] to forecast the maximum and the minimum temperature for seven days, given weather data for the past two days. The linear regression model outperformed the functional regression model.

Finally, in [21], artificial neural networks (ANN) were developed using TensorFlow in order to predict the temperature. Four cases have been studied in both experiments, using prediction horizons of 1, 3, 6 and 12 hours.

III. Materials and Methods

A. Dataset Description

Our research was based on a dataset from the Kaggle [22]. This dataset includes data found in the period 1/1/2012 - 31/12/2015. We have 1460 instances (one per day). All the attributes (5 as input to ML models and 1 for target class) are described as follows:

- **date** (DD-MM-YYYY): This feature captures the date of the measurement.
- **precipitation** (mm) [23]: This feature captures the total rainfall depth per day, expressed in millimeters. It's numeric data.
- **temp max** (°C) [24]: This feature captures the value of the maximum temperature per day, expressed in Celsius. It's numeric data.
- **temp min** (°C) [24]: This feature captures the value of the minimum temperature per day, expressed in Celsius. It's numeric data.
- **wind** (Beaufort Scale) [25]: This feature captures the average daily value of the wind force, expressed in the Beaufort Scale. It's numeric data.
- **weather**: This feature captures the weather conditions. It consists of 5 classes (3.63% drizzle, 43.87% rain, 43.81% sun, 1.78% snow, 6.91% fog). It's nominal data.

B. Multi-class Classification Methodology

In this research paper, the weather forecast is approached as a multi-class classification task. In order to approach such problems, the one-against-all and one-against-one methods are used in the relevant literature [26], [27]. We assume K classes, which correspond to the K weather conditions in the dataset under consideration. In the present work, there are $K = 5$ classes that capture weather conditions: drizzle, rain, sun, snow and fog.

The one-against-one (OAO) method splits the multi-class classification problem into $K = 5$ binary classification sub-problems. Thus, the initial dataset is divided into one dataset for each of the $K = 5$ classes versus every other individual class. Unlike (OAO), the one-against-all (OAA) constructs $K = 5$ models where the j-th model of these, is trained with all of the subjects in the j-th class with positive labels and the rest of subjects with negative labels. These schemes are outlined in Table I. In order to solve the binary classification sub-problems, LR [28] and SVM [19] models will be exploited.

TABLE I
MULTI-CLASS CLASSIFICATION SCHEMES OAO VS OAA FOR WEATHER FORECASTING $K = 5$

OAA			**OAO**					
drizzle	rain sun snow fog	*1 binary classifier*	**snow**	rain	sun	fog	drizzle	*K-1 binary classifiers*
sun	rain snow drizzle fog	*1 binary classifier*	**sun**	rain	-	fog	drizzle	*K-2 binary classifiers*
rain	drizzle snow fog sun	*1 binary classifier*	**drizzle**	rain	-	fog	-	*K-3 binary classifiers*
fog	drizzle rain sun snow	*1 binary classifier*	**fog**	rain	-	-	-	*K-4 binary classifiers*
snow	drizzle sun rain fog	*1 binary classifier*	-	-	-	-	-	-
Total		*K binary problems*	**Total**	-	-	-	-	*K(K-1)/2 binary problems*

C. Data Preprocessing and Features Analysis

For the purpose of this study, raw data were preprocessed (by combining random oversampling to the minority class and

TABLE II
STATISTICAL CHARACTERISTICS

Features	Min	Max	Mean ± std
precipitation	0	55.9	2.913 ± 6.141
temp max	-1.6	35	14.224 ± 8.080
temp min	-7.1	18.3	6.404 ± 5.784
wind	0.4	8.8	3.113 ± 1.462

random undersampling to the majority class [29]) such that the instances in all classes are uniformly distributed and the trained models are unbiased.

Table II presents the statistical characteristics of the numerical features in the balanced dataset. We observe that the mean precipitation is 3.029 mm. The mean value of the maximum (temp max) and minimum temperature (temp min) is 16.439 and 8.235 °C, respectively. The minimum value of the precipitation is 0 mm, and the maximum value is 55.9 mm. The lowest and highest value of maximum temperature is -1.6°Cand 35.6°C, correspondingly. The lowest and highest value of the minimum temperature is -7.1°Cand 18.3°C. Also, the mean wind speed is 3.113. Finally, the minimum value of the wind speed is 0.4, and the maximum value is 9.5 Beaufort.

IV. RESULTS AND DISCUSSION

In this section, we will present the results of the experiments following the strategies OAO and OAA. The performance of these approaches is compared to single classifiers k-NN with k = 1 and Random Forest. The optimal settings for the considered models are presented in Table III. The experiments were executed in the WEKA environment by exploiting the libsvm library. To evaluate the performance of the investigated models and methods, Accuracy, Recall, F-Measure and AUC are captured under 10-fold cross-validation. The definition of these metrics is captured in Table IV.

TABLE III
OPTIMAL PARAMETERS SETTING

Optimal parameters	
Classifiers	**Hyper-parameters**
SVM (Linear Kernel)	Default hyperparameters
SVM (polynomial kernel)	C = 1, coef0 =30, degree = 3 gamma = 0.1
Logistic Regression	Default hyperparameters
Random Forest	maxDepth = 1 numFeatures = 2
k-NN	Default hyperparameters k = 1

Tables V, VI show the performance of SVM under the two multi-class classification strategies. In each method, we investigated for which kernel function the model is more efficient. Focusing on OAO, the linear SVM exhibited significant superiority in classification achieving an Accuracy of 96.64%, average values of Precision, Recall and F-Measure very close to accuracy and AUC of 98.5%. In terms of the OAA strategy, polynomial SVM of degree = 3 was the most efficient. Comparing the average outcomes of the best models in each strategy, the experiment highlighted the former better than the latter by $\sim 3.2\%$ in terms of Precision, Recall and F-Measure, 3.58% concerning Accuracy and with similar AUC values.

TABLE IV
PERFORMANCE METRICS DEFINITION

Precision	$\frac{1}{K}\sum_{i=1}^{K}\frac{TP_i}{TP_i+FP_i}$
Recall	$\frac{1}{K}\sum_{i=1}^{K}\frac{TP_i}{TP_i+FN_i}$
F-Measure	$2\frac{Precision \cdot Recall}{Precision+Recall}$
Accuracy	$\frac{1}{K}\sum_{i=1}^{K}\frac{TN_i+TP_i}{TN_i+TP_i+FN_i+FP_i}$

TABLE V
PERFORMANCE OF OAO WITH LINEAR SVM

OAO - linear SVM - Accuracy = 96.64%				
Precision	**Recall**	**F-Measure**	**AUC**	**Class**
1.000	0.997	0.998	1.000	**drizzle**
1.000	0.908	0.952	0.958	**rain**
0.902	0.976	0.938	0.981	**sun**
0.977	1.000	0.988	0.996	**snow**
0.962	0.952	0.957	0.989	**fog**
0.968	0.966	0.966	0.985	**Average**

TABLE VI
PERFORMANCE OF OAA WITH 3RD-DEGREE POLYNOMIAL SVM

OAA - polynomial $d = 3$ SVM - Accuracy = 93.08%				
Precision	**Recall**	**F-Measure**	**AUC**	**Class**
0.831	0.997	0.907	0.998	**drizzle**
0.955	0.877	0.914	0.968	**rain**
0.914	0.836	0.873	0.984	**sun**
0.980	1.000	0.990	0.997	**snow**
0.996	0.945	0.970	0.979	**fog**
0.935	0.931	0.935	0.985	**Average**

Tables VII, VIII show the performance of LR under the two multi-class classification strategies. It is observed that under LR, which is employed to solve the binary classification problem, the OAO strategy is again more efficient by 0.5 to 4.1% in terms of Precision, Recall and F-Measure and 3.23% in terms of Accuracy.

Tables IX, X show the performance of k-NN and Random Forest classifiers which are suitable for multi-class classification problems and thus employed for comparison to the SVM and LR executed under strategies OAO and OAA. Figure 1 illustrates the average performance of all models. Comparing Random Forest to OAO linear SVM, the latter remains the most competent model in all metrics and a candidate solution for the topic under consideration.

In the above tables of SVM, perfect recall values are observed for the snow class. But the precision is perfect for the drizzle and rain classes only in the case of OAO. Also, the perfect AUC value is obtained only for the drizzle class. In the LR model, there is a perfect precision value for all classes except drizzle and rain which attained low values impacting the average precision of the OAA LR. In the case of OAO LR, only the recall of the snow class reached the upper limit

TABLE VII
PERFORMANCE OF OAA WITH LR

OAA - LR - Accuracy = 83.69%				
Precision	**Recall**	**F-Measure**	**AUC**	**Class**
0.735	0.997	0.846	0.998	**drizzle**
0.654	0.860	0.743	0.937	**rain**
1.000	0.384	0.554	0.784	**sun**
1.000	1.000	1.000	1.000	**snow**
1.000	0.945	0.972	0.984	**fog**
0.878	0.837	0.823	0.941	**Average**

TABLE VIII
PERFORMANCE OF OAO WITH LR

OAO - LR - Accuracy = 86.92%				
Precision	**Recall**	**F-Measure**	**AUC**	**Class**
0.762	0.997	0.864	0.977	**drizzle**
0.978	0.620	0.759	0.867	**rain**
0.848	0.781	0.813	0.894	**sun**
0.893	1.000	0.943	0.988	**snow**
0.933	0.949	0.941	0.983	**fog**
0.883	0.869	0.864	0.942	**Average**

TABLE IX
PERFORMANCE OF 1-NN MODEL

1-NN - Accuracy = 89.04%				
Precision	**Recall**	**F-Measure**	**AUC**	**Class**
0.890	0.997	0.940	0.985	**drizzle**
0.887	0.777	0.828	0.918	**rain**
0.837	0.723	0.776	0.886	**sun**
0.980	1.000	0.990	0.997	**snow**
0.853	0.955	0.901	0.967	**fog**
0.889	0.890	0.887	0.951	**Average**

TABLE X
PERFORMANCE OF RANDOM FOREST MODEL

Random Forest - Accuracy = 92.60%				
Precision	**Recall**	**F-Measure**	**AUC**	**Class**
0.912	0.997	0.953	0.999	**drizzle**
1.000	0.777	0.875	0.979	**rain**
0.920	0.908	0.914	0.987	**sun**
0.849	1.000	0.918	1.000	**snow**
0.982	0.949	0.965	0.995	**fog**
0.933	0.926	0.925	0.992	**Average**

(namely, one). Similar results concerning recall were produced by 1-NN and Random Forest. Also, the latter achieved perfect precision only in the rain class. Focusing on the AUC, the more close to one is its value the more appropriate the model will be in the correct distinction between the five weather categories. In other words, the AUC values show that the explored models with high probability will be able to separate the instances into the five classes. With an in-depth search of the SVM kernel type [30] at each multi-class classification scheme, we see that linear SVM with OAO is a competitive approach against OAA. Similar conclusions are derived for the RF model. Considering the average performance metrics outcomes, we conclude that linear SVM is the most efficient model.

V. CONCLUSIONS

AI and ML constitute important tools in weather forecasting (expected weather conditions) besides various other applications such as water quality prediction, sentiment analysis,

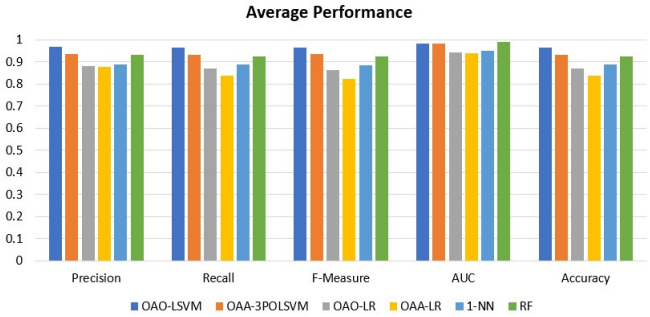

Fig. 1. Average performance of all considered models

chronic condition risk prediction, fraud detection, etc. This study describes a multi-class framework investigating the performance of Logistic Regression and Support Vector Machine under two strategies and comparing them to two commonly used (and efficient) classifiers Random Forest and k-NN. All models presented high AUCs indicating the promising separation ability concerning the five classes of weather conditions. However, the SVM classifier under the OAO strategy achieved more concise outcomes and generally superior performance than the rest ones.

In future work, we aim to explore richer information data and the capabilities provided by other multi-class classification models and/or schemes from the Machine, and Deep Learning point of view [31].

ACKNOWLEDGMENT

This research was funded by the European Union and Greece (Partnership Agreement for the Development Framework 2014-2020) under the Regional Operational Programme Ionian Islands 2014-2020, project title: “Indirect costs for the project ‘TRaditional corfU Music PresErvation through digiTal innovation’ ”, project number: 5030952.

REFERENCES

[1] V. Krishnamurthy, “Predictability of weather and climate,” *Earth and Space Science*, vol. 6, no. 7, pp. 1043–1056, 2019.

[2] T. Sharp, “Earth’s atmosphere: Composition, climate & weather,” *Space. com. Recuperado de¿ https://www. space. com/17683-earth-atmosphere. html*, 2017.

[3] K. L. Ebi, J. Vanos, J. W. Baldwin, J. E. Bell, D. M. Hondula, N. A. Errett, K. Hayes, C. E. Reid, S. Saha, J. Spector *et al.*, “Extreme weather and climate change: population health and health system implications,” *Annual review of public health*, vol. 42, p. 293, 2021.

[4] A. Brazienė, J. Venclovienė, V. Vaičiulis, D. Lukšienė, A. Tamošiūnas, I. Milvidaitė, R. Radišauskas, and M. Bobak, “Relationship between depressive symptoms and weather conditions,” *International journal of environmental research and public health*, vol. 19, no. 9, p. 5069, 2022.

[5] V. Spiridonov and M. Ćurić, *Fundamentals of Meteorology*. Springer, 2021.

[6] K. A. Pattani and S. Gautam, “Introduction to meteorology and weather forecasting,” in *Artificial Intelligence of Things for Weather Forecasting and Climatic Behavioral Analysis*. IGI Global, 2022, pp. 1–15.

[7] S. Murugan Bhagavathi, A. Thavasimuthu, A. Murugesan, C. P. L. George Rajendran, L. Raja, and R. Thavasimuthu, “Weather forecasting and prediction using hybrid c5. 0 machine learning algorithm,” *International Journal of Communication Systems*, vol. 34, no. 10, p. e4805, 2021.

[8] H. Friesen, *Meteorologist*. Weigl Publishers, 2019.

[9] A. H. Murphy, "Probabilistic weather forecasting," in *Probability, statistics, and decision making in the atmospheric sciences*. CRC Press, 2019, pp. 337–377.

[10] É. P. Caillault, A. Bigand *et al.*, "Comparative study on univariate forecasting methods for meteorological time series," in *2018 26th European Signal Processing Conference (EUSIPCO)*. IEEE, 2018, pp. 2380–2384.

[11] D. Mauree, D. S.-H. Lee, E. Naboni, S. Coccolo, and J.-L. Scartezzini, "Localized meteorological variables influence at the early design stage," *Energy Procedia*, vol. 122, pp. 325–330, 2017.

[12] N. Singh, S. Chaturvedi, and S. Akhter, "Weather forecasting using machine learning algorithm," in *2019 International Conference on Signal Processing and Communication (ICSC)*. IEEE, 2019, pp. 171–174.

[13] Y. Liu, J.-W. Bi, and Z.-P. Fan, "A method for multi-class sentiment classification based on an improved one-vs-one (ovo) strategy and the support vector machine (svm) algorithm," *Information Sciences*, vol. 394, pp. 38–52, 2017.

[14] E. Dritsas, S. Alexiou, and K. Moustakas, "Copd severity prediction in elderly with ml techniques," in *Proceedings of the 15th International Conference on PErvasive Technologies Related to Assistive Environments*, 2022, pp. 185–189.

[15] R. Mantri, K. R. Raghavendra, H. Puri, J. Chaudhary, and K. Bingi, "Weather prediction and classification using neural networks and k-nearest neighbors," in *2021 8th International Conference on Smart Computing and Communications (ICSCC)*. IEEE, 2021, pp. 263–268.

[16] A. Pandey, C. Agrawal, and M. Agrawal, "A hadoop based weather prediction model for classification of weather data," in *2017 second international conference on electrical, computer and communication technologies (ICECCT)*. IEEE, 2017, pp. 1–5.

[17] A. Kanavos, M. Trigka, E. Dritsas, G. Vonitsanos, and P. Mylonas, "A regularization-based big data framework for winter precipitation forecasting on streaming data," *Electronics*, vol. 10, no. 16, p. 1872, 2021.

[18] R. I. Rasel, N. Sultana, and P. Meesad, "An application of data mining and machine learning for weather forecasting," in *International conference on computing and information technology*. Springer, 2017, pp. 169–178.

[19] K. Purwandari, J. W. Sigalingging, T. W. Cenggoro, and B. Pardamean, "Multi-class weather forecasting from twitter using machine learning aprroaches," *Procedia Computer Science*, vol. 179, pp. 47–54, 2021.

[20] M. Holmstrom, D. Liu, and C. Vo, "Machine learning applied to weather forecasting," *Meteorol. Appl*, pp. 1–5, 2016.

[21] E. B. Abrahamsen, O. M. Brastein, and B. Lie, "Machine learning in python for weather forecast based on freely available weather data," 2018.

[22] "Weather forecasting," https://www.kaggle.com/datasets/ananthr1/weather-prediction, (accessed on 15 September 2022).

[23] B. Atkinson, "Precipitation," in *Man and Environmental Processes*. Routledge, 2019, pp. 23–37.

[24] E. Gravanis, E. Akylas, and G. Livadiotis, "Physical meaning of temperature in superstatistics," *EPL (Europhysics Letters)*, vol. 130, no. 3, p. 30005, 2020.

[25] M. H. Davis, "A beaufort scale of predictability," in *The Fascination of Probability, Statistics and their Applications*. Springer, 2016, pp. 419–434.

[26] M. Awad and R. Khanna, "Support vector machines for classification," in *Efficient learning machines*. Springer, 2015, pp. 39–66.

[27] A. A. Kurniawan, K. Usman, and R. Y. N. Fuadah, "Classification of tropical cyclone intensity on satellite infrared imagery using svm method," in *2019 IEEE Asia Pacific Conference on Wireless and Mobile (APWiMob)*. IEEE, 2019, pp. 69–73.

[28] S.-H. Moon and Y.-H. Kim, "An improved forecast of precipitation type using correlation-based feature selection and multinomial logistic regression," *Atmospheric Research*, vol. 240, p. 104928, 2020.

[29] R. Mohammed, J. Rawashdeh, and M. Abdullah, "Machine learning with oversampling and undersampling techniques: overview study and experimental results," in *2020 11th international conference on information and communication systems (ICICS)*. IEEE, 2020, pp. 243–248.

[30] W. Shu and K. Cai, "A svm multi-class image classification method based on de and knn in smart city management," *IEEE Access*, vol. 7, pp. 132 775–132 785, 2019.

[31] Q. Yang, L. Tan, B.-Q. Wu, G.-L. Tian, L. Xu, J.-T. Yang, J.-H. Jiang, and R.-Q. Yu, "Beyond one-against-all (oaa) and one-against-one (oao): An exhaustive and parallel half-against-half (hah) strategy for multi-class classification and applications to metabolomics," *Chemometrics and Intelligent Laboratory Systems*, vol. 204, p. 104107, 2020.