Lossy Predictive Models for Accurate Classification Algorithms

Aekyeung Moon
ETRI
akmoon@etri.re.kr

Seung Woo Son
University of Massachusetts Lowell
seungwoo_son@uml.edu

Hyson Kim
ETRI
etri2022_khs@etri.re.kr

Minjun Kim
Andong National University
manjun1004@student.anu.ac.kr

2022 IEEE International Conference on Big Data (Big Data) | 978-1-6654-8045-1/22/$31.00 ©2022 IEEE | DOI: 10.1109/BigData55660.2022.10020381

***Abstract*—Recent years have witnessed an upsurge of interest in lossy compression due to its potential to significantly reduce data volume with adequate exploitation of the spatiotemporal properties of IoT datasets. However, striking a balance between compression ratios and data fidelity is challenging, particularly when losing data fidelity impacts downstream data analytics noticeably. In this paper, we propose a lossy prediction model dealing with binary classification analytics tasks to minimize the impact of the error introduced due to lossy compression. We specifically focus on five classification algorithms for frost prediction in agricultural fields allowing preparation by the predictive advisories to provide helpful information for timely services. While our experimental evaluations reaffirm the nature of lossy compressions where allowing higher errors offers higher compression ratios, we also observe that the classification performance in terms of accuracy and F-1 score differs among all the algorithms we evaluated. Specifically, random forest is the best lossy prediction model for classifying frost. Lastly, we show the robustness of the lossy prediction model based on the data fidelity in prediction performance.**

***Index Terms*—Classification, Transform Coding, Lossy Compression, Data Augmentation, IoT**

I. Introduction

With the recent report by the Food and Agriculture Organization (FAO) of the United Nations that climate change could reduce crop yields by up to 30% by 2050 [1], scientists have increasingly focused on predicting climate phenomena known to have a substantial impact on agriculture [2], [3]. Among various climate phenomena affecting the agricultural sector, frost or freeze damage to flowers and buds at or near the bloom stage in spring could result in significant crop failures [4], [5].

The convergence of AI (artificial intelligence) and IoT (internet of things) in the agriculture domain, when applied effectively exploiting data extracted from a steady stream of raw data by sensors, afford decision-supporting methods through actionable prediction knowledge discovery [3]. As the risk of late-spring frosts increases, actionable predictive knowledge can be helpful for farmers to protect plants and farms effectively. For example, Rozante et al. [2] have studied the damage caused by the frost phenomenon and the prevention through prediction. Chung et al. [6] showed that an accurate frost forecast would minimize frost damage by taking preventive actions. These studies demonstrate that more intelligent agriculture promises predictive insights using the data adequately captured in the agriculture domain to enable proactive steps.

While agricultural services like frost prediction through this predictive knowledge look promising, providing accurate services in resource-constrained edge devices is challenging due to highly imbalanced labels and the need for quantifying the impact of lossy data compression on classification algorithms. This paper presents a lossy prediction model for the analytics task of agricultural domains to discover climate conditions and extract meaningful knowledge from environmental IoT datasets applied through lossy compressions. Furthermore, to manage IoT environmental data efficiently and reliably, we build the prediction model in a lossless way while running inference tasks using lossy data. We then evaluate the fidelity of the reconstructed data using DCT (discrete cosine transform)-based lossy compression algorithms to assess the impact of the lossy compression and restoration on the performance of classification algorithms. We choose DCT because of its high decorrelation efficiency [7], its inverse has the same spectrum as the nearly original data, and a low error rate between original and reconstructed datasets [3], [8].

We conduct extensive evaluations of our prediction models based on five machine learning models popularly used in classification tasks to predict frost. Our results demonstrate that Random Forests (RF) present superior performances in accuracy and F1-score among five lossy models we evaluated [9]: Decision Tree (DT), Random Forest (RF), Adaptive Boosting (AdaBoost), Support Vector Machine (SVM), and Artificial Neural Networks (ANN). We also optimize RF using the stratified k-fold cross-validation method and verify that the prediction performances by the optimized model improve accuracy, precision, recall, and F1-score between the original and reconstructed data from lossy compression. Lastly, our results demonstrate that fixed information-based lossy compression reduces the required data storage while maintaining data quality sufficient for accurate classification performance.

II. Preliminaries

A. Classification Task for Frost Prediction

Due to climate changes, the occurrence of frost during late spring (off-seasonal) has recently increased. This late spring frost could worsen crop damage because flowers have feeble freezing resistance compared to the dormant period, as deviations in low-temperature increase after flowering [5], [10]. Therefore, a precise prediction of frost can minimize crop damage by allowing preventive measures (passive protection)

or actions during the frost, e.g., moving a frost fan around the crop [11].

Several studies have shown that predictive models using various IoT datasets could forecast frost effectively. For example, Lee et al. [12] attempted to use Logistic Regression (LR) and DT techniques to predict frost using frost observation data. In contrast, Kim [13] estimated the occurrence of frost using ANN, RF, and SVM. However, these studies needed to verify the proposed models with field observations (i.e., ground truth); more importantly, they could suffer from class imbalance problems. More recently, Noh et al. [5] used five input data variables (i.e., temperature, subzero temperature duration, precipitation, wind speed, and humidity) and considered the imbalanced class problem. Specifically, they applied the under-sampling technique to solve imbalanced class and the degree of affecting frost depending on the time to notify the preparation to prevent frost. They also adjusted the ratio of frost events to 50:50 by applying the under-sampling method but with a possibility of valuable data loss.

The frost prediction is part of the predictive services we have been providing since 2015 using microclimate datasets collected from IoT weather stations[1]. Moreover, we have collected observed frost data by farmers since 2017 through the deployed system because we need datasets with labels to develop a prediction model. The frost prediction model can use these labeled datasets to classify between non-frost and frost states in a training phase. However, the quantity of our labeled datasets is still limited; thus, we need methods to augment data for highly accurate prediction.

B. Lossy Compression

As the amount of data produced by IoT weather stations increases continuously (with more deployed nodes and sensors), resource-scarce devices frequently employ data reduction techniques to lessen the volume of data and overhead by exploiting potential *redundancy* in spatial, temporal or both [14]. Like several recent studies, we utilize lossy compression using a signal transformation to exploit characteristics in many time-series datasets [8]. Lossy compression could filter noises and transfer potentially vital information to analytic workloads running on the cloud because bandwidth, energy, and storage constrain IoT stations considerably. Therefore, we need to manage these data efficiently to reduce storage and transmission costs [15]. While employing lossy data compression can be helpful for highly efficient data management, it still needs a comprehensive evaluation that establishes the criteria for selecting the degree of compromise in data quality and makes reasonably accurate forecast services.

Several prior studies demonstrated that lossy compression could help reduce the data size, while error rates and loss of data quality are often hard to bound [16], [17]. However, as reported in several prior studies, such as analyses of electrocardiogram (ECG) data [18], weather data [3], and high-performance computing applications [19], data reconstructed from lossy compression allows for meaningful analysis. Nevertheless, lossy compression techniques are still subjective to data fidelity issues for the analytics of prediction tasks because acceptable information loss varies [3], [20]. No studies have quantified such investigation on analytical tasks.

Among many lossy compression techniques, transform-based lossy algorithms utilize spatial data characteristics better. There are two considerations to exploit transform-based lossy compression effectively. First, most lossy compression techniques capitalize that their overall patterns are spatiotemporally smooth in time-series datasets [8], [21]. Because of this, data compression in conjunction with transformation techniques can be more effective, and the transformed data usually explicitly reveal the information's correlation. Second, the distortion by loss data is minor, thus confirming that one can reconstruct IoT data (e.g., temperature) by maintaining minimal fractions of the original data.

The goal of our lossy compressor is to find k, which is the number of significant coefficients in the transformed domain required to store outcomes L_k of original datasets X, such that we can reconstruct data within a certain tolerance. We then characterize the k needed to approximate X and the reconstructed $R(L_k)$ from L_k. The rate-distortion of L_k measures $ER_k|X, R(L_k)|$. $ER_k|X, R(L_k)|$ means the error rate between original datasets X and reconstructed datasets ($R(L_k)$) from lossy compressed datasets (L_k). Then we analyze the impact of $ER_k|X, R(L_k)|$ for the achievable classification performance using our prediction model for accurate classification.

III. Datasets Preparation for Highly Accurate Prediction Task

In this section, we describe datasets to build the lossy prediction model for frost classification and steps to increase the model's predictability. As depicted in Fig. 1, the overall process involves learning (or training) and on-time prediction phases. In the learning phase, we collect observation data with labels and augment data for high-accuracy prediction. To manage IoT datasets efficiently (less burden on data management) and reliably (less information loss on prediction model), we apply lossy compression, store lossy data, and then reconstruct them for the prediction task.

A. Observed Datasets

Depending on the weather condition on the farm at a particular time, we predict when adverse weather conditions (i.e., frost) could occur. It would ultimately provide an instant alert to prevent or minimize damage from frost by allowing farmers to act proactively. In this paper, we consider the environmental data in an orchard region in South Korea, where we have been collecting data using the deployed IoT stations since 2015. Datasets mainly include the microclimate datasets collected from IoT stations and the local weather data obtained from KMA (Korea Meteorological Administration). These real-world IoT datasets are continuously monitored and collected every minute from the IoT stations. Specifically,

[1]http://183.106.117.219

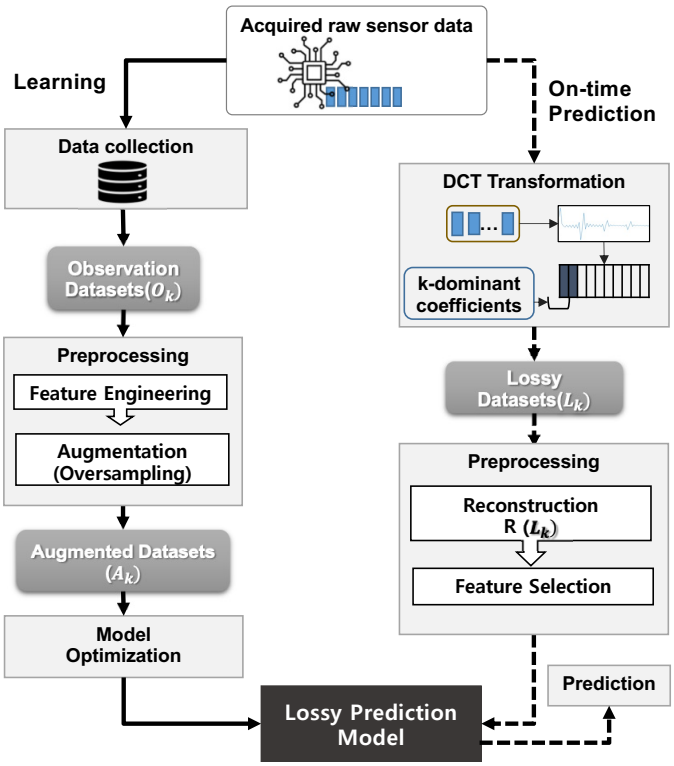


Fig. 1: Overall process of our Lossy Prediction Model.

the environmental data collected are air temperature, grass temperature (i.e., temperature above grass), humidity, wind speed, rainfall, and soil moisture. We also extract the most relevant features to predict accurately from these collected datasets. The extracted features contain dew point, temperature inversion, rainfall, solar radiation, minimum air temperature, temperature difference, wind speed, minimum grass temperature, and soil moisture.

After selecting features, we labeled the datasets with farmer's frost observation to evaluate the prediction model. However, one of the main obstacles to using observational data, infrequent but critical events like frost, for learning a model is imbalanced datasets. Even though IoT stations can provide rich environmental datasets for machine learning, it is impossible to anticipate evenly labeled training data. Table I shows an imbalance ratio between the minority (frost) and the majority (no-frost), obtained through observation of frost. This imbalanced condition leads to biased outcomes by the majority class. In other words, most machine learning-based algorithms can be overwhelmed by the majority of data unless they adequately address the imbalance problem. Considering that the given class is balanced when the imbalance ratio (majority class/minority class) is close 0, almost all stations suffer from the severe class imbalance problem except station H. For instance, even though station H's imbalance ratio is much better than others (4.04 vs. greater than 43 for stations A through G), its data also might negatively influence machine learning-based classifiers because it produced poor prediction

TABLE I: Data distribution of observation datasets that shows a class imbalance.

Station	# of valid data	# of majority	# of minority	Imbalance ratio
A	1124	1104	20	55.20
B	1124	1099	25	43.96
C	1124	1114	10	111.40
D	1121	1102	19	58.00
E	1124	1110	14	79.29
F	1122	1109	13	85.31
G	2139	2119	20	105.95
H	1944	1558	386	4.04

performance with only 0.2632 in F1-score.

B. Augmented Datasets

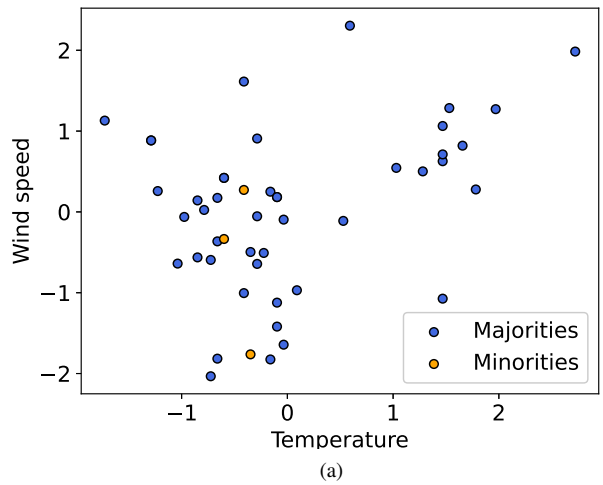
Since our frost prediction is a binary classification task, we can use supervised machine learning-based predictive models. One critical precondition for effectively training a supervised machine learning model is obtaining large-scale datasets with proper labels [22]. In other words, the learning datasets need a balanced ratio between classification classes. Therefore, we synthetically augment to alleviate the class imbalance issue discussed in Section III-A. Therefore, we employ the Synthetic Minority Oversampling Technique (SMOTE) method [23], which strengthens the pros and makes up for the cons of the oversampling technique, to solve the class imbalance problem in our training data for frost prediction.

Fig. 2 illustrates our mechanism using original and augmented datasets. SMOTE augments the minority of the original datasets in Fig. 2a as shown in Fig. 2b. Its fundamental principle is to generate synthetic minority samples on the new random position in the feature space by using k-nearest neighbors and uniform random variables to lessen the influences of the overfitting problem.

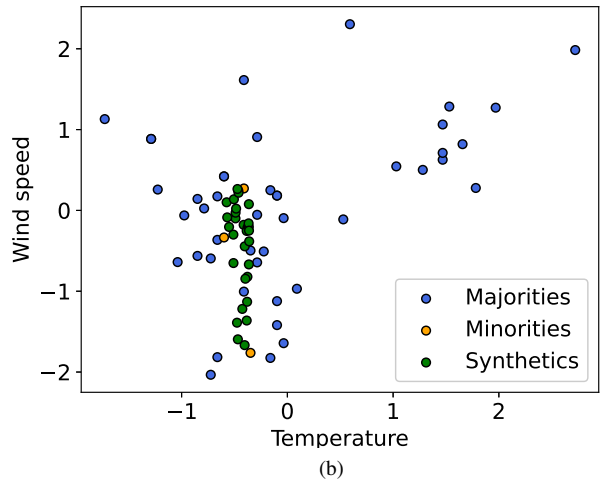
C. Lossy Datasets

While we use raw datasets with data augmentation to build a robust classification model, the prediction uses lossy datasets. It is noteworthy that any form of additional measurements, such as more diverse sensors, an increased number of weather stations or sensor nodes, or even sensing frequencies, is intended to build a model that predicts more accurately. On the other hand, such increased data collection will significantly burden IoT devices where data collection and inference occur. Our design of prediction utilizes the lossy model using minimizing data for performing prediction at IoT devices.

As our compression utilizes transform methods, we perform data transforms, specifically DCT. Once transformed in the frequency domain, we obtain the relationship between the percentage of informative DCT coefficients (i.e., low-frequency ones) and the amount of energy those coefficients carry. Due to DCT's high compaction property, a few low-frequency DCT components retain the most energy (i.e., information), and the remaining high-frequency coefficients are close to zero. Our prior studies also indicate that maintaining coefficients containing 99.9% of energy (i.e., information) and discarding



(a)



(b)

Fig. 2: Two representative sampling methods: (a) original, (b) augmented datasets.

insignificant coefficients K provides competitive compression ratios. However, the implication of such lossy compressions on the commonly-used ML prediction model applied to the reconstructed $R(X_k)$, i.e., how much information is still meaningful without reducing prediction accuracy, is largely unexplored.

IV. LPM: Lossy Prediction Model Methodology

This section presents Lossy Prediction Model (LPM) design, as illustrated in Fig. 1. The LPM consists of learning (including the evaluation model) and accurate prediction. To build an accurate prediction Model, in learning phase, we uses augmented observational dataset and uses lossy datasets in on-time prediction to impose less burden on data management and prediction model.

A. Lossy Prediction Model

To describe how our lossy prediction model works, let us consider the data point $X_{(t)}^{(i)} \in \mathbb{R}$ in Equation 1, where $1 \leq t \leq n$, and n is the number of types of sensor data. In $1 \leq i \leq p$, i means frost prediction during specific period. It also requires a labeled class, $\theta^{(i)}$, which characterizes frost conditions for each day. Let $\theta^{(i)}$ denote the parameter associated with typical frost characteristics of fn feature selected in period i (one day). We augment minority data $\{A_{(1)}^{(i)}, A_{(t)}^{(i)} ..., A_{(fn)}^{(i)}, \theta^{(i)}\}$ to alleviate the class imbalance issue.

$$X_{(t)}^{(i)} = \begin{Bmatrix} X_{(1)}^{(1)}, X_{(2)}^{(1)} ..., X_{(n)}^{(1)} \\ X_{(1)}^{(2)}, X_{(2)}^{(2)} ..., X_{(n)}^{(2)} \\ ... \\ X_{(t)}^{(p)}, X_{(2)}^{(p)} ..., X_{(n)}^{(p)} \end{Bmatrix} \Rightarrow \left\{ A_{(1)}^{(i)}, A_{(2)}^{(i)} ..., A_{(fn)}^{(i)}, \theta^{(i)} \right\}. \tag{1}$$

To apply lossy datasets (L_k) in the predictive model Fig. 1, lossy compression for $X_{(t)}^{(i)}$ follows the condition in Equation 2. We define $\hat{X}_t$, which denotes transformed components for a compressed block size of N at a given period t: $\hat{X}_t = \{\hat{X}_{t,1}, \hat{X}_{t,2}, ...\hat{X}_{t,N}\}$. Thus, each coefficient component has its own energy coefficient defined as $e(\hat{X}_{t,n})$. $EC(\hat{X}_{t,k})$ is formulated as the energy concentration (EC) contained in the number of coefficients components, denoted as k, of the entire transformed components ($\hat{X}_t$), which is calculated as:

$$L_k = EC(\hat{X}_{t,k}) = \frac{\sum_{n=1}^{k} e(\hat{X}_{t,n})^2}{\sum_{n=1}^{N} e(\hat{X}_{t,n})^2} > \delta, n = 1, 2, ..., N, k \leq n. \tag{2}$$

It is noteworthy that k determines how many coefficients are required to represent δ amount of the energy [3]. In other words, k refers to the number of dominant coefficients representing block $\hat{X}_t$. EC means the difference between original datasets X and reconstructed datasets $R(L_k)$ from L_k. The goal of LPM is assumed as Equation 3 even $ER_k|X, R(L_k)|$ exists.

$$LPM(X_{(t)}^{(i)}) \approx LPM[R(EC(\hat{X}_{t,k}))] = LPM[R(L_k)]. \tag{3}$$

B. Agricultural Feature Engineering

Features in Table II define $X_{(t)}^{(i)}$ in Equation 1 to process LPM for frost forecast. Table II lists essential environmental data for LPM, such as temperature and relative humidity used for training our prediction model. We provide a forecast at 23:00 (or 11 p.m.) so that farmers can receive predictive advisory of frost information about the next day and have sufficient preparation time to prevent damage. Lastly, the significance of some features is time sensitive, mostly from noon to our forecast time (23:00), as described in the calculation model in Table II.

Additional features, such as temperature inversion and dew point, are calculated from raw datasets using Equations 4-7. The temperature inversion, a reversal of the expected temperature behavior, means that a layer of warmer air overlies

that of cool air at the surface. As a result, frost will be highly likely to happen if the temperature inversion occurs. To obtain temperature inversion, we use two types of thermometer sensor data to capture the inversion: one collected from 10cm above (T_{grass}) the ground and the other from 1.5m above the ground (T_{air}). We then calculate a temperature inversion layer using those two temperatures. The temperature inversion denoted as *I* is calculated as:

$$\Delta T_t = T_{grass}(t) - T_{air}(t),\ T_s \leq t \leq T_e, \tag{4}$$

$$I(\Delta T_t) = \sum \Delta T_t, \quad \text{if } \Delta T_t < 0, \tag{5}$$

where t is a measurement time, and the difference between the grass and air temperatures is ΔT_t.

The dew point, another essential feature affecting frost conditions, refers to the temperature at which water vapor in the atmosphere is saturated, and part of the water vapor condenses with water. Above this temperature, the moisture will stay in the air. A well-known approximation used to calculate the dew point denoted as T_{dp}, given the actual air temperature, T_{air} (in degrees Celsius), and relative humidity (in percent), RH, is calculated as follows:

$$\gamma(T_{air}, RH) = ln(\frac{RH}{100}) + \frac{b * T_{air}}{c + T_{air}}, \tag{6}$$

$$T_{dp} = \frac{c * \gamma(T_{air}, RH)}{b - \gamma(T_{air}, RH)}, \tag{7}$$

where b = 17.62 and c = 243.5, one of the constant sets used in National Oceanic and Atmospheric Administration and other studies.

C. Evaluating Lossy Prediction Models

To quantify which classification algorithm would work best as a lossy prediction, we evaluate the prediction model of five representative classification methods, DT, RF, AdaBoost, SVM, and ANN.

Once we validate the best-performing prediction model with lossy datasets, we optimize the selected model using the k-fold cross-validation method. The k-fold cross-validation method divides a given dataset into k non-overlapping folds. The k-th fold uses a validation set and the remainder as a training set. We select optimized hyperparameters by deriving and comparing averaged performance by combining the grid search method. The stratified k-fold cross-validation method mitigates the influence of remained class imbalance problem. We derive the optimized classifier for our lossy prediction model from these processes.

We sorted out and utilized the same test data in the modeling process for a fair comparison. Similarly, z-score normalization uses hyperparameters in the modeling and the missing value process. After the optimized classifier derives experimental results, we evaluate and compare the results based on the confusion matrix. Since our test data is imbalanced, we consider both accuracy and F1-score performance criteria.

V. Results and Discussion

A. Evaluation Metrics

We use the following metrics to evaluate the performance of the lossy prediction model: compression ratio (CR) for measuring the amount of data reduced, the Peak Signal-to-Noise Ratio (PSNR) for measuring information loss, and various other metrics for measuring the impact of LPM on prediction.

- Compression Ratio (CR): The compression ratio is given by $CR = \frac{|D|-|D'|}{|D|} \times 100\%$, where $|D|$ is the size of D, $|D'|$ is the reduced size, and δ is the amount of energy.
- Let X be the original data and lossy data L_k be the reconstructed data. Then, we measure the peak signal-to-noise ratio (PSNR) for $ER_k|X, R(L_k)|$, a commonly used average error metric.

$$\text{PSNR} = 20log_{10}(\frac{Max(x) - Min(x)}{RMSE(x, R(L_k))}.$$

- The objective of our lossy prediction model is to achieve a higher Accuracy and F1-score. TP, TN, FP, and FN indicate true positive, true negative, false positive, and false negative, respectively.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{FN} + \text{TN}}.$$

$$\text{F1-score} = \frac{\text{TP}}{\text{TP} + 0.5 \times (\text{FP} + \text{FN})}.$$

B. Evaluation of Lossy Datasets

Since our predictive services run on lossy data, let us first discuss how much data reduction our compression mechanism brings. In our evaluation, we compress fixed energy (or information) compaction rate δ = 0.99 (99%) in Equation 2 and reconstruct the five data variables (Temperature, Grass, Soil Moisture, Wind Speed, Relative Humidity), each with 525,190 data points.

Table III shows each data's CR, when δ = 0.99 (99%), and data characteristics in terms of the normalized standard deviation (NSTD), skewness, and kurtosis. We observe that each data shows similar characteristics in all metrics. In the case of skewness, wind speed has a positive value. Wind speed has a higher NSTD than that of the others. In the case of the temperature, CR is 99.35, which means 0.65% of data is needed, whereas wind speed needs 39.1% of data is needed.

To measure how much the reconstructed data deviate from the original data, i.e., $ER_k|X, R(L_k)|$, we use PSNR to measure distortion estimates. A higher value of PSNR represents lower error and is similar to the original datasets. Overall, the effect of the error rate varies depending on the datasets. For instance, the error rate for wind speed increases more than others. Soil moisture data shows higher PSNR than the other datasets, which means that $R(L_k)$ is closer to the original.

C. Evaluation of Lossy Prediction Performance

We next evaluate the performance of our frost prediction task using lossy datasets. We utilize machine learning libraries from scikit-learn [24] to train and evaluate five machine

TABLE II: The environmental datasets used in training.

	features	calculation model
Temperature Inversion	I(ΔT_t)	Equation 4 and 5
Dew Point	T_{dp}	Equation 6 and 7
Rainfall	$rain$	between 22:30 and 23:00
Solar Radiation	$solar$	between 12:00 and 19:00
Air Temperature	$Tair(min)$	the minimum temperature between 12:00 and 23:00
	$Tair(max)$	the maximum temperature between 12:00 and 23:00
	$Tair(diff)$	T(max) - T(min)
Wind Speed	$Wspeed$	between 21:00 and 23:00
Grass Temperature	$Tgrass$	between 12:00 and 23:00
	$Tgrass(min)$	minimum between 12:00 and 23:00
Soil Moisture	$Tsoil$	between 22:30 and 23:00

TABLE III: The evaluated datasets and their characteristics.

	NSTD	Skewness	Kurtosis	CR	PSNR
Temperature	0.81	-0.072	-0.94	99.35	42.03
Grass Temperature	0.91	-0.047	-0.85	98.62	41.87
Soil Moisture	0.25	-0.22	-1.034	99.99	51.2
Wind Speed	1.097	1.17	1.25	60.9	37.32
Relative Humidity	0.31	-0.61	-0.74	99.93	33.07

TABLE IV: Experimental results for model selection.

Station	Performance	RF	SVM	DT	ANN	AdaBoost
A	Accuracy	**0.9132**	0.8852	0.8663	0.8935	0.8663
	F1-score	**0.6908**	0.6494	0.5944	0.6532	0.5858
B	Accuracy	**0.8959**	0.8508	0.8632	0.8508	0.8632
	F1-score	**0.6712**	0.5720	0.6562	0.5947	0.6575
C	Accuracy	**0.9162**	0.8867	0.8844	0.8958	0.8829
	F1-score	**0.6337**	0.5609	0.5905	0.6203	0.5953
D	Accuracy	**0.8784**	0.8343	0.8406	0.8407	0.8527
	F1-score	**0.6545**	0.6053	0.6032	0.6042	0.6322
E	Accuracy	**0.9335**	0.8869	0.9066	0.8711	0.9003
	F1-score	**0.6344**	0.1936	0.6190	0.1598	0.6175
F	Accuracy	**0.9135**	0.8762	0.8876	0.8975	0.8937
	F1-score	**0.6365**	0.5521	0.6103	0.6195	0.6224
G	Accuracy	**0.9919**	0.9081	0.9880	0.9214	0.9876
	F1-score	**0.9538**	0.0441	0.9290	0.3590	0.9264
H	Accuracy	**0.9786**	0.7548	0.9580	0.7235	0.9506
	F1-score	**0.9743**	0.7132	0.9479	0.6641	0.9384

learning algorithms: DT, RF, AdaBoost, SVM, and ANN. We set 70% of the data for training and the remaining 30% for the test. Also, we augment synthetic minority samples 2x for station H and 10x for the remaining stations to make a similar imbalance ratio across all stations.

Table IV shows the experimental results for the five classification algorithms we tested. As we can see, RF shows the best performance among all five models in both Accuracy and F1 score. We attribute RF's superior performance on lossy datasets as follows. RF reduces variance by combining various subtrees based on operating principles, although it might slightly increase bias. The variance reduction is often meaningful for the model's overall performance improvement, guaranteeing robust results. Based on these results for model selection, we conclude that RF works best to work with reconstructed data from lossy compression. We next perform hyperparameter optimization for RF. Specific hyperparameters we obtained are as follows:

- the number of subtrees (from 10 to 200)
- the maximum depth of each subtree (from 5 to 30)
- the splitting criteria (Gini index and Entropy)

VI. Conclusion

The convergence of IoT and AI has the potential to produce predictive advisories like frost forecasts using a large volume of diverse IoT data, which needs to be stored efficiently and with maximum data fidelity. In this paper, we evaluated the effectiveness of the lossy prediction model on IoT environmental datasets as an exemplar of the predictive service. Our experimental results show that lossy compressions based on DCT can achieve significantly higher compression ratios with a marginal loss of data quality. Specifically, we compared the performance of frost prediction tasks using RF, SVM, DT, ANN, and AdaBoost on the reconstructed data using various error-bounding methods to evaluate the feasibility of applying lossy compressions. Moreover, our LPM (Lossy Prediction Model) has an accuracy of 0.88 – 0.99 with 91.8% of the average compression ratio, demonstrating that a certain degree of loss in data fidelity by lossy compression hardly affects the accuracy of frost prediction outcomes.

Acknowledgements

This work was supported by Electronics and Telecommunications Research Institute (ETRI) grant funded by the Korean government (No. 22YD1100) and the Andong-si support project (No. 22AD1100). This material is also, in part, based upon work supported by the National Science Foundation under Grant No. 1751143. The Titan X Pascal used for this research was donated by the NVIDIA Corporation. In addition, this research was supported by the MSIT (Ministry of Science, ICT), Korea, under the National Program for Excellence in SW), supervised by the IITP (Institute of Information & communications Technology Planing & Evaluation) in 2019 (2019-0-01113).

References

[1] R. Chandra and S. Collis, "Digital Agriculture for Small-Scale Producers: challenges and Opportunities," *Communications of the ACM*, vol. 64, no. 12, 2021.

[2] J. R. Rozante, E. R. Gutierrez, P. L. da Silva Dias, A. de Almeida Fernandes, D. S. Alvim, and V. M. Silva, "Development of an index for frost prediction: Technique and validation," *Meteorological Applications*, 2019.

[3] A. Moon, J. Kim, J. Zhang, and S. W. Son, "Evaluating Fidelity of Lossy Compression on Spatiotemporal Data from an IoT Enabled Smart Farm," *Computers and Electronics in Agriculture*, vol. 154, pp. 304–313, Nov. 2018.

[4] P. Matzneller, K.-P. Götz, and F.-M. Chmielewski, "Spring frost vulnerability of sweet cherries under controlled conditions," *International Journal of Biometeorology*, vol. 60, no. 1, pp. 123–130, 2016.

[5] I. Noh, H.-W. Doh, S.-O. Kim, S.-H. Kim, S. Shin, and S.-J. Lee, *Machine Learning-Based Hourly Frost-Prediction System Optimized for Orchards Using Automatic Weather Station and Digital Camera Image Data*. Atmosphere, 2021.

[6] U. Chung, H. C. Seo, and J. I. Yun, "Site-Specific Frost Warning Based on Topoclimatic Estimation of Daily Minimum Temperature," *Korean Journal of Agricultural and Forest Meteorology*, vol. 6, no. 3, pp. 164–169, 2004.

[7] A. Moon, J. Kim, J. Zhang, and S. W. Son, "Lossy Compression on IoT Big Data by Exploiting Spatiotemporal Correlation," in *2017 IEEE HPEC*, 2017, pp. 1–7.

[8] J. Zhang, X. Zhuo, A. Moon, H. Liu, and S. W. Son, "Efficient Encoding and Reconstruction of HPC Datasets for Checkpoint/Restart," in *Symposium on Mass Storage Systems and Technologies (MSST)*, 2019.

[9] P. Möller-Acuña, R. Ahumada-García, and J. A. Reyes-Suárez, "Machine learning for prediction of frost episodes in the maule region of chile," in *International conference on ubiquitous computing and ambient intelligence*. Springer, 2017, pp. 715–720.

[10] K. Young-A, "The spatial distribution and recent trend of frost occurrence days in South Korea," *Journal of The Korean Geographical Society*, vol. 41, 2006.

[11] R. Snyder and J. de Melo-Abreu, "Frost protection: Fundamentals, practice and economics," *Rome: Food and Agriculture Organization of the United Nations*, 2005.

[12] H. Lee, J. Chun, H. Han, and S. Kim, "Prediction of Frost Occurrences Using Statistical Modeling Approaches," in *Hindawi Publishing Corporation Advances in Meteorology*, 2016.

[13] Y. Kim, K.-M. Shim, M.-P. Jung, and I.-T. Choi, "Study on the Estimation of Frost Occurrence Classification Using Machine Learning Methods," in *Korean Journal of Agricultural and Forest Meteorology*, 2017.

[14] J. Zhang, A. Moon, X. Zhuo, and S. W. Son, "Towards Improving Rate-Distortion Performance of Transform-Based Lossy Compression for HPC Datasets ," in *IEEE HPEC*, 2019.

[15] T. Bose, S. Bandyopadhyay, S. Kumar, A. Bhattacharyya, and A. Pal, "Signal Characteristics on Sensor Data Compression in IoT – An Investigation," in *13th Annual IEEE International Conference on Sensing, Communication, and Networking (SECON)*, June 2016, pp. 1–6.

[16] A. Moon, X. Zhuo, J. Zhang, and S. W. Son, "AD^2: Improving Quality of IoT Data through Compressive Anomaly Detection," in *2019 IEEE Big Data*, 2019.

[17] J. Zhang, J. Chen, X. Zhuo, A. Moon, and S. W. Son, "DPZ: Improving Lossy Compression Ratio with Information Retrieval on Scientific Datas," in *IEEE International Conference on Cluster Computing*, 2021.

[18] A. Moon, S. W. Son, J. Jung, and Y. J. Song, "Understanding Bit-Error Trade-off of Transform-based Lossy Compression on Electrocardiogram Signals," in *Proceedings of IEEE Big Data*, 2020, pp. 3494–3499.

[19] K. Zhao, S. Di, X. Liang, S. Li, D. Tao, J. Bessac, Z. Chen, and F. Cappello, "SDRBench: Scientific Data Reduction Benchmark for Lossy Compressors," in *IEEE International Conference on Big Data (Big Data)*, 2020, pp. 2716–2724.

[20] S. Li, N. Marsaglia, V. Chen, C. Sewell, J. Clyne, and H. Childs, "Performance Impacts of In Situ Wavelet Compression on Scientific Simulations," in *Proceedings of the In Situ Infrastructures on Enabling Extreme-Scale Analysis and Visualization*, 2017, pp. 37–41.

[21] A. Moon, X. Zhuo, J. Zhang, S. W. Son, and Y. J. Song, "Anomaly Detection in Edge Nodes using Sparsity Profile," in *Proceedings of IEEE Big Data*, 2020, pp. 1236–1245.

[22] H. Ren, B. Xu, C. Y. Yujing Wang, C. Huang, X. Kou, T. Xing, M. Yang, J. Tong, and Q. Zhang, "Time-Series Anomaly Detection Service at Microsoft ," in *KDD*, 2019.

[23] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.

[24] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.