SMOTE and Gaussian Noise Based Sensor Data Augmentation

Mehmet Arslan
Gazi University
Department of Computer Engineering
Ankara,Turkey
mehmetarslaan07@gmail.com

Metehan Guzel
Gazi University
Department of Computer Engineering
Ankara,Turkey
metehanguzel@gazi.edu.tr

Mehmet Demirci
Gazi University
Department of Computer Engineering
Ankara,Turkey
mdemirci@gazi.edu.tr

Suat Ozdemir
Gazi University
Department of Computer Engineering
Ankara,Turkey
suatozdemir@gazi.edu.tr

***Abstract*— Deep learning achieves successful prediction results by training multilayer neural network based machine learning models on large amounts of data. One of the best ways to improve performance of artificial neural networks is to add more data to the training set. In the literature, some data augmentation techniques have been developed for this purpose and they are widely used in processing image data. For example, data can be enriched by replicating the images in the training set with views from different angles and sufficient data can be obtained to find a generalizable model. In this study, we focused on augmenting sensor data by applying image data augmentation methods. A data set including temperature, humidity, light, and air quality sensor data was augmented using two different data augmentation techniques (SMOTE Regression and Gaussian Noise), and their effect on the performance of an LSTM model which estimates missing or incorrect values was investigated. RMSE values obtained by using real and augmented data were compared to evaluate the impact of both techniques. The most successful test estimation model from the data set was the air quality model. In addition, it was concluded that SMOTE regression gave better results when the two techniques were compared.**

Keywords—Deep Learning, Neural Networks, Data Augmentation, Smote Regression, Gaussian Noise, LSTM

I. INTRODUCTION

Deep learning has gained great success in many applications such as image analysis, speech recognition and text comprehension as one of the most remarkable machine learning techniques. In classification and pattern recognition applications, supervised and uncontrolled strategies are used to learn multilevel features in hierarchical architectures. It is clear that data volumes need to be increased in order to give better results to deep learning models that are complicated by the development of applications. In deep learning approaches, it is known that the more data available, the better the model will be learned [1]. In deep learning and other neural networks, it is tried to solve the problems such as the learning of the models enough or class imbalance by different methods. In particular, data augmentation is a common practice for increasing the size of the training data set and is also used as an editing technique that makes the model more resistant to slight changes in input data. Data augmentation is a technique that allows artificially creating new training data from existing training data. In machine learning applications such as image recognition or video processing, data augmentation is usually performed by applying small transformations or noise to existing samples in the dataset [2,3]. In addition, the use of data augmentation in natural language processing is quite common. Since changing the order of the word also changes the meaning of the word, augmenting the data in natural language processing is quite different from that in the image data [4]. In scientific studies, it is seen that data augmentation techniques are generally applied in text and visual data. In general, the main purpose of data augmentation is to see possible variations of the samples in the training set by the model. This study investigates the feasibility of the data augmentation for other data types.

Recent developments in sensor networks and communication technologies have facilitated the collection of large data. In daily life, deep learning and neural networks are used to develop solutions against many problems using sensor data. In this study, the data collected to measure the quality of air through sensors are analyzed and data augmentation techniques are applied to these data. Machine learning and statistical data analysis methods are used during data augmentation. Our contributions can be summarized as follows:

- A LSTM based indoor air quality prediction method is proposed;
- Application of data augmentation to enhance accuracy of LSTM based prediction model is proposed and two augmentation methods, namely SMOTE and Gaussian Noise are employed for augmentation;
- Through experimental study on original data, the potential of the proposal is shown.

The rest of the paper is organized as follows. We first introduce the related works in Section II. Then, we give essential information about LSTM, Gaussian Noise, SMOTE Regression and RMSE in Section III. We present the LSTM based prediction models in Section IV. The experiments and experimental results is given in Section V. The paper is summarized and concluded in Section VI.

This work was supported by the Scientific and Technological Research Council of Turkey (TÜBİTAK), Grant No: 118E212

978-1-7281-3964-7/19/$31.00 ©2019 IEEE

II. Related Works

Convolutional neural networks have performed well on large data sets in recent years. For CNNs, it is difficult to apply only small data sets**.** In addition, one of the difficulties encountered in studies with medical imaging is that models require larger training data. Collection of medical data is a costly task and needs to be done in collaboration with the field experts. To overcome this difficulty, a Generative Adversarial Networks (GAN) based data augmentation method for medical images is proposed [5]. The aim of the study is to increase classification accuracy in medical imaging data sets with limited number of samples. GAN is applied to the data set to increase the diversity of the data and extensive tests are conducted. Results indicate meaningful increase in both sensitivity and specificity [5].

In another study [6], application of data augmentation in image recognition is investigated. Random Deletion (RD) method is employed for augmentation. RD creates new samples by deleting pixel values at randomly selected regions of the images. CIFAR10, CIFAR100 and Fashion-MNIST are selected as data sets and a convolutional neural network (CNN) type of model is used for recognition. Numerous CNN based models with different architectures are trained with the augmented data sets. Test results indicate improved recognition performance and high interoperability with CNN [6].

It is often difficult to collect and label large amounts of sensor data. In this article, researchers developed Parkinson's Disease tracking algorithm based on wearable sensor data. In this research, they increased the performance of the algorithm by applying various augmentation techniques to the sensor data. The graphs obtained with the collected data constitute the inputs of the CNN model. Data augmentation techniques were applied to these images. Recommended augmentation methods are jittering, scaling, rotating, permutation, magnitude-warping and time warping. PD monitoring and suggested augmentation methods were successfully tackle using a 7-layer CNN. It was seen that the samples augmented by applying the combination of rotating and permutation data augmentation techniques increased the performance of the model from 77.52% to 86.88% [7].

In another study on augmenting data in sensor data, performance tests were performed on Deep LSTM artificial neural network using data augmentation on data collected from wearable sensors monitoring human activity. The researchers suggested using spectral properties to learn human activity. The algorithm steps that are suggested as a method of increasing data are as follows;

- Spectral features data [set 1]
- Apply local averaging on set 1: G=4
- Combine down-sampled data with the spectral features dataset to get [set2]
- Shuffle set 2 randomly
- Apply local averaging on set 2: G = 2
- Combine local averaged data with set 2 to get [set3]

The feature extraction and augmentation algorithm proposed in this study provided improved learning accuracy and performed close to the non-augmented data set [8].

III. Methodology

In this section, we will talk about the techniques used in the data augmentation model and LSTM, which is the neural network model that we use to observe the effect of the augment of data. The RMSE metric used to test the data we have trained in our LSTM model and to measure the changing error with the actual data set will be described in this section.

A. LSTM (Long-Short Term Memory)

Long short-term memory (LSTM) is a repetitive neural network (RNN) architecture, remembering values at random intervals. LSTMs protect the memory cell from errors and eliminate the problem of missing parameters (lost gradient problem) between processes in neural network [9]. It allows learning from arrays of hundreds of time intervals. Stored values are not changed when progress is learned in a learned manner. RNNs allow back and forth connections between neurons. Feedback loops are what allows repeating networks to be better than other neural networks in pattern recognition

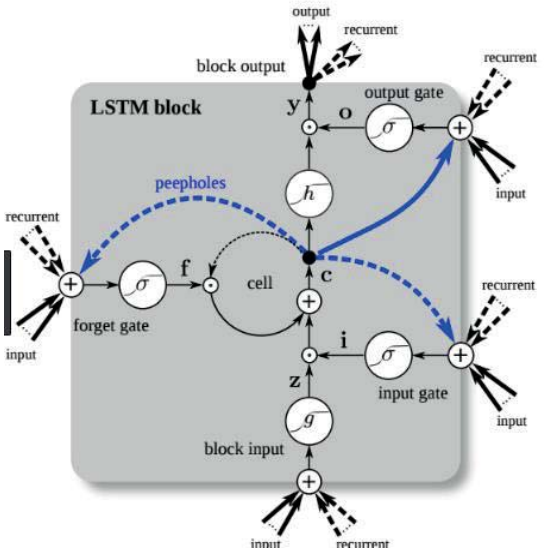

Fig. 1. A Long Short-Term Memory Unit Model [10].

Fig. 1. The LSTM unit has an entry and exit port that provides information flow to and from the inside of the forget gate and unit

Block Input: $z^t = g\,(W_z\,X^t + R_z\,y^{t-1} + b_z)$
Input Gate: $i^t = \sigma\,(W_i\,X^t + R_i\,y^{t-1} + p_i\,\Theta\,c^{t-1} + b_i)$
Forget Gate: $f^t = \sigma\,(W_f\,X^t + R_f\,y^{t-1} + p_f\,\Theta\,c^{t-1} + b_f)$
Cell State: $c^t = i^t\,\Theta\,z^t + f^t\,\Theta\,c^{t-1}$
Output Gate: $o^t = \sigma\,(W_o\,X^t + R_o\,y^{t-1} + p_o\,\Theta\,ct + b_o)$
Block Output: $y^t = o^t\,\Theta\,h\,(c^t)$

When we look at the above equations, it is seen that the equations of input, forget and exit doors are the same and only the matrices calculated in these equations are different. The activation function in the forget gate in Fig. 1 is defined as the sigmoid function in each LSTM block. Activation functions in other gates may be changeable. The weights W_z, W_i, W_f, W_o at time t are R_z, R_i, R_f, R_o are repetitive weights. *C* is a hidden state calculated based on the current entry and the previous hidden (candidate) state. In a sense the internal memory of the *C* block. b_z, b_i, b_f, b_o values are bias vectors. A simple LSTM model has only one hidden LSTM layer, while an advanced LSTM layer model contains multiple hidden layers. LSTM has been applied to a variety of real problems such as machine translation, speech recognition, neuro linguistic model, and image recognition.

In this study, models with recurrent neural network LSTM will be developed for time series prediction problem.

B. Gaussian Noise

Gaussian noise is statistical noise with the function of equal probability density in the normal distribution, also known as Gaussian distribution. Normal distribution is the most important and most widely used distribution in statistics. Sometimes called the "bell curve". The noise produced here is random, but like most random events, a certain pattern follows. The reason why Gaussian distribution is highly preferred in data science or machine learning is that Gauss random events are very common in nature. When we observe a random event that is the sum of many independent events, all random variables appear to be Gauss variables [11].

Gaussian noise to a data set can be added as follows:

1. Random noise is calculated and assign noise to the variable.
2. Add the noise to the data set (Data set = Data set + Noise)

The Gaussian distribution probability density function is calculated by the following formula.

$$N(\chi:\mu,\sigma)=\frac{1}{\sqrt{2\pi\sigma^2}}\exp(-\frac{1}{2}(x-\mu)^2/\sigma^2) \qquad (1)$$

Using this technique, noise was generated in a random manner to augmented the data set.

C. SMOTE Regression

While developing prediction models, sometimes data sets with unstable class variables are encountered. SMOTE (Synthetic Minority Over Sampling Technique) is used to illustrate the minority situations for these class imbalance datasets. Based on the given ratio, this algorithm produces synthetic samples using the nearest neighbors. SMOTE is a common technique for augmenting data for classification problems [12,13]. A system in which regression and data augmentation are used together is demonstrated by this proposed technique. Depending on the amount of over-sampling required, the nearest neighbors are randomly selected to produce synthetic data from the actual data set. In this paper, the nearest five neighbors are used.

D. Root Mean Squared Error (RMSE)

It is a metric that measures the magnitude of the error that is most preferred in finding the distance between the values that a machine learning or deep learning model predicts as a result of the trainings and the actual values.
The RMSE metric is calculated by the following formula;

$$\sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j-y_j^{'})^2} \qquad (2)$$

Where y_j is the real value and $\hat{y}_j$ is the estimated value in Eq. 2. By using RMSE, a difference in the error between the data and the actual data was observed by comparing the model with different data sets. The reason we choose this metric is that RMSE is the default metric of many models, although it is prejudiced against more complex and higher deviations. The loss function defined in terms of RMSE can be uniquely differentiated and facilitates mathematical operations. In addition, LSTM time series estimation models have been found to be better with RMSE metric [14].

IV. Problem Implementation

The fact that deep learning requires big data for a quality result is the main feature that distinguishes it from standard machine learning algorithms. As a solution to this problem, it has been focused on providing deep learning algorithms to provide better results with augmented data by meeting data need with data augmentation techniques. The algorithms developed in the project and the estimation models that will be used together with these algorithms were carried out as planned. First of all, we use a training and a test set because we want to estimate the error our system will actually have. Since the data of the test set should be as close to the real data as possible, the data augmentation process was applied only to the training set.

A. Dataset

The data set consists of temperature, humidity, light, IAQ (In Door Air Quality), Sharp in, Sharp Out, Sharp Difference and PIR values collected from the sensors where air quality is measured. The data set contains data collected from a closed environment at 5 minute intervals. In the developed models, the features which are seen as more determinative in the data set are used by selecting them. The data set we used contains 4925 samples and features temperature, humidity, light and air quality.

TABLE I. INPUT AND OUTPUT OF THE MODELS

Models	Input I	Input II	Input III	Input IV	Output
Mdl 1	$Temp_{t-1}$	Hum_{t-1}	Lux_{t-1}	IAQ_{t-1}	IAQ_t
Mdl 2	Hum_{t-1}	Lux_{t-1}	IAQ_{t-1}	$Temp_{t-1}$	$Temp_t$
Mdl 3	Lux_{t-1}	IAQ_{t-1}	$Temp_{t-1}$	Hum_{t-1}	Hum_t
Mdl 4	IAQ_{t-1}	$Temp_{t-1}$	Hum_{t-1}	Lux_{t-1}	Lux_t

Table 1 shows the input and output values of the models. At the moment t-1, the values calculated for the four features are the inputs of the model and the value of the feature to be estimated at the moment of t is the output of the models.

TABLE II.STATISTICS OF THE DATASET

	Features			
Statistics	Temperature	Humidity	Lux	IAQ
Count	4925.0	4925.0	4925.0	4925.0
Min	0.000	0.000	0.000	0.000
Max	33.412	47.825	429.667	194.111
Mean	19.480	31.240	54.349	80.506
Std.	4.094	5.893	91.718	24.965

These statistics will later shed light on the evaluation of model performances in Section VI.

V. Test Results

In four different LSTM models in which the temperature, humidity, light and air quality characteristics are estimated one by one the test error of the augmented data on the models and the test error of the non- augmented data

were compared. The proposed Gaussian Noise method randomly determines the noise to be added. Therefore, the data set was applied to the models closest to the real data set.

TABLE III . STATISTICS GAUSSIAN NOISE AND AUGMENTED DATASET

	Features			
Statistics	Temperature	Humidity	Lux	IAQ
Count	9850.0	9850.0	9850.0	9850.0
Min	-2.599	-2.726	-3.391	0.000
Max	34.736	48.792	431.070	194.111
Mean	19.485	31.217	54.347	80.506
Std.	4.158	5.931	91.710	24.963

TABLE IV. STATISTICS SMOTE REGRESSION AND AUGMENTED DATASET

	Features			
Statistics	Temperature	Humidity	Lux	IAQ
Count	9850.0	9850.0	9850.0	9850.0
Min	0.000	0.000	0.000	0.000
Max	33.413	47.825	429.667	194.111
Mean	19.470	31.253	54.312	80.487
Std.	4.045	5.838	91.647	24.943

When the statistics of the normal data shown in Table 1 are compared with the 2-fold augment of the data set in Table II and Table III, the min, max, mean and standard deviation values are closer in the SMOTE Regression technique.

TABLE V TEST RMSE RESULTS FOR ALL PREDICTION MODELS

Data Augmentation Metrics	**Test error (RMSE)**			
	Temperature	Humidity	Lux	IAQ
Normal Data	2.477	2.452	1.440	0.311
Gaussian Noise	**2.458**	2.774	1.626	**0.203**
SMOTE Regression	2.474	2.782	1.612	**0.130**

The table above shows the results of test errors of normal data and augmented data. These results obtained by augmenting the data set by 2-fold. In SMOTE regression, the number of neighbors was taken as 5 and the standard deviation of the noise generated in the other technique as 1.

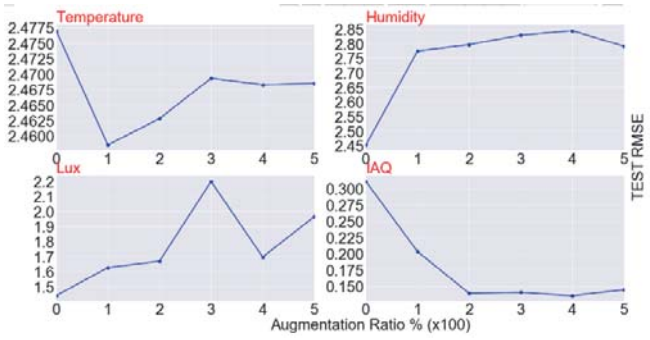

Fig. 4. Test RMSE changes according to the augmentation rate by Gaussian Noise technique in all prediction models. (x-axis shows the augmentation ratio and y-axis test RMSE values)

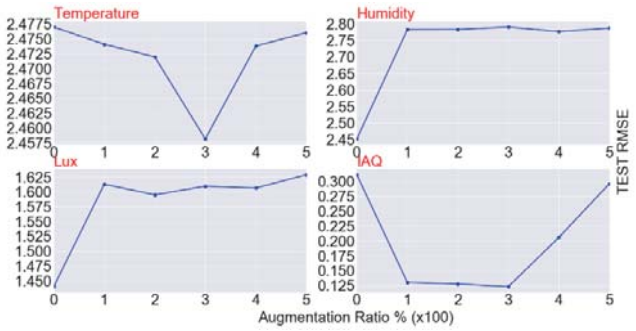

Fig. 5. Test RMSE changes according to the augmentation rate by SMOTE Regression technique in all prediction models. (x-axis shows the augmentation ratio and y-axis test RMSE values)

As shown in Fig. 4 and Fig. 5, when the data is augmented using these two methods, it is seen that the test error decreases in temperature and air quality (IAQ) estimation models. These graphs show the error of the test in this increment by augmenting the data 5-fold starting from the normal data. SMOTE regression showed some increases in errors for IAQ and temperature, but all results were measured below normal data.

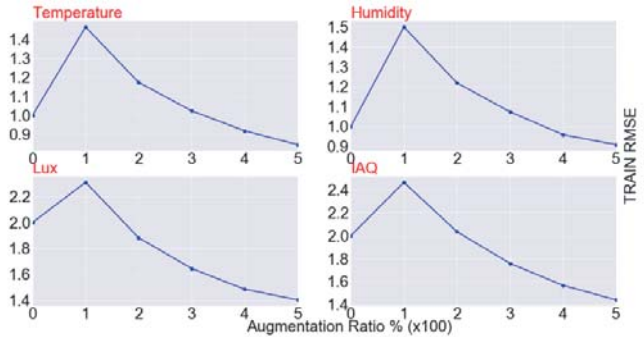

Fig. 6. Train RMSE changes according to the augmentation rate by SMOTE Regression technique in all prediction models. (x-axis shows augmentation ratio and y-axis train RMSE values)

Fig. 6 shows a decrease in the training errors measured for SMOTE Regression technique in all models.

When the results of our study are evaluated together with the experiments, various data augmentation techniques can be applied for numerical data and will contribute to the development of machine learning and deep learning models. In this study, augmenting with Gaussian Noise has some disadvantages compared to SMOTE Regression method. In the IAQ prediction model, the 0.311 test error measured with normal data was reduced to 0.130 by the SMOTE regression method. In addition, train errors were measured using this method and it is seen that the general course of train errors decreased for all models. Since Gaussian noise is a statistical method, it uses randomly generated noise according to the standard deviation determined when generating noise, resulting in negative values in the data set. Therefore, the data set augmented by this technique is not sufficiently similar to the normal data set.

As a result, it is clearly seen that SMOTE regression is better than Gaussian noise. It was observed that the data sets obtained by increasing the augmentation ratio in SMOTE regression to is largely close to the normal data set. Thus, it is shown in Fig. 6 that the training error increased for the first augmentation ratio for each model but decreased as the rate increased. Augmenting the data reduced the error rate of air quality and temperature prediction models. In general,

the error rate performed close to the non-increased data set with an improved learning rate with the proposed techniques.

VI. Conclusion and Future Directions

In this work, we proposed two algorithms to augmenting the sensor data, namely SMOTE Regression and Gaussian Noise based algorithms. Using the data set, models are trained and test procedures are conducted. The results of the two methods are compared by testing the augmented data in the trained models with the real data set. When the test results are compared with the proposed algorithms, it is observed that the RMSE error value decreases in IAQ and temperature models. In addition, the augmented data obtained by using Smote Regression showed a good performance when we look at the measured educational results.

In this study, data pre-processing step is taken into consideration in order to have positive results of models other than air quality and temperature. In addition, MCMC (Markov Chain Monte Carlo) algorithm as a different data augmentation technique on digital data is another method that we have encountered in the literature [15]. We can recommend the MCMC method as well as the techniques we use for the researchers who will work on this subject. This algorithm can be developed with the library named Pym3 in Python programming language and data augmentation can be realized by making inferences on real data set models with probabilistic programming.

References

[1] H. Chiroma et al., "Progress on Artificial Neural Networks for Big Data Analytics: A Survey," IEEE Access, pp. 1-1, 2018.

[2] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks %J Commun. ACM," vol. 60, no. 6, pp. 84-90, 2017.

[3] L. Perez and J. Wang, "The Effectiveness of Data Augmentation in Image Classification using Deep Learning," CoRR, vol. abs/1712.04621, / 2017.

[4] X. Lu, B. Zheng, A. Velivelli, and C. Zhai, "Enhancing text categorization with semantic-enriched representation and training data augmentation," (in eng), J Am Med Inform Assoc, vol. 13, no. 5, pp. 526-535, Sep-Oct 2006.

[5] M. Frid-Adar, E. Klang, M. Amitai, J. Goldberger and H. Greenspan, "Synthetic data augmentation using GAN for improved liver lesion classification," 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), Washington, DC, 2018, pp. 289-293.

[6] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, "Random Erasing Data Augmentation," *CoRR,* vol. abs/1708.04896, / 2017.

[7] T. T. Um et al., "Data Augmentation of Wearable Sensor Data for Parkinson's Disease Monitoring using Convolutional Neural Networks," CoRR, vol. abs/1706.00527, / 2017.

[8] O. Steven Eyobu and D. S. Han, "Feature Representation and Data Augmentation for Human Activity Classification Based on Wearable IMU Sensor Data Using a Deep LSTM Neural Network," (in eng), Sensors (Basel), vol. 18, no. 9, p. 2892, 2018.

[9] S. Squartini, A. Hussain, and F. Piazza, "Preprocessing based solution for the vanishing gradient problem in recurrent neural networks," 2003. [Online]. Available: https://doi.org/10.1109/ISCAS.2003.1206412.

[10] K. Greff, R. K. Srivastava, J. Koutník, B. R. Steunebrink, and J. Schmidhuber, "LSTM: A Search Space Odyssey," IEEE Trans. Neural Netw. Learning Syst., vol. 28, no. 10, pp. 2222-2232, / 2017.

[11] C. E. Rasmussen and C. K. I. Williams, Gaussian processes for machine learning (Adaptive computation and machine learning). MIT Press, 2006, pp. I-XVIII.

[12] S. C. Wong, A. Gatt, V. Stamatescu, and M. D. McDonnell, "Understanding Data Augmentation for Classification: When to Warp?," 2016. [Online]. Available: https://doi.org/10.1109/DICTA.2016.7797091.

[13] L. Torgo, R. P. Ribeiro, B. Pfahringer, and P. Branco, "SMOTE for Regression," 2013. [Online]. Available: https://doi.org/10.1007/978-3-642-40669-0_33.

[14] S. Bouktif, A. Fiaz, A. Ouni, and A. M. Serhani, "Optimal Deep Learning LSTM Model for Electric Load Forecasting using Feature Selection and Genetic Algorithm: Comparison with Machine Learning Approaches †," Energies, vol. 11, no. 7, 2018.

[15] J. P. Hobert, The data augmentation algorithm: Theory and methodology. 2011.