Data Augmentation on Limited Biometric Data Set for M2M Authentication Model Testing

Rin Nadia*, Dana Koshen †, JaeSeung Song*
Department of Computer and Information Security, Sejong University, Seoul, Republic of Korea*
Department of Computer Science, Sejong University, Seoul, Republic of Korea †
E-mail: rinnadia@sju.ac.kr; danakosh@sju.ac.kr; jssong@sejong.ac.kr (Corresponding author)

***Abstract*—Examining the performance of artificial intelligence (AI)-based model for machine to machine (M2M) authentication needs a large number of data. The accessibility of open biometric data is restricted by privacy regulations such as general data protection regulation (GDPR). This is because the data is commonly obtained by sensors embedded in personal wearable devices. Thus exploring and developing a system with biometric data as the main parameter is difficult to do. As data augmentation is a technique to increase the size of the data set by producing derived data from the original data, its usage in AI is popular especially for computer vision. Incorporating data augmentation in the development of AI-based authentication model could solve the shortage of database. Therefore this study proposes a data augmentation model and analyzes its training and augmenting performance.**

***Index Terms*—Deep Learning, Data Augmentation, IoT, Time Series**

I. Introduction

Biometric data provides a good characteristic for various researches ranging from medical studies to security studies. While current research trends involving biometric data typically includes AI elements inside it. The development of AI models whether it is categorized as machine learning model or deep learning model ideally requires a large data set. Because training a model with a small dataset can cause the over fitting problem. As biometric data is labeled as confidential and sensitive data, in the recent decades, its accessibility and availability to public is heavily regulated. This limitation could hinder the development of M2M biometric-based authentication from having further progress. Some studies like [1] show that the publicly available time series biometric datasets have a small number of data.

Data augmentation is concept to acquire another data based on original data elements. In computer vision, data augmentation is usually conducted by several techniques such as position augmentation and color augmentation. Both of them require no pattern learning. However the time series biometric data is patterned data by its nature. The usual augmentation techniques applied in image data could not be utilized for expanding time series biometric data. This means that the augmentation toward time series dataset requires different approaches compared to image data.

To summarize, this paper works on analyzing data augmentation experiment toward human walking pattern as an alternative to solve a small dataset problem in M2M authentication model development. This paper is planned as follows. Section II is for the related works. Section III explains the experiment preparation and the augmentation and evaluation methodology is in section IV. Section V is for result and augmentation analysis and section VI concludes the paper.

II. Related Work

Biometric data, as popular as it is, is challenging to obtain to be sufficient for use in a learning process of different algorithms. This issue is especially important during the times when big data is emerging as an essential area of computer science. Because of the small size of the open data sets on biometric data, researchers are learning new ways of expanding the available volume of data. The problem of insufficient data for training deep learning models is not exclusive to biometric data acquisition but is a general challenge that concerns other types of data sets. The article was written on different ways data augmentation can be performed on the primary data set while preserving the original data distribution [2]. The data used by the authors to explain and describe different data augmentation processes was of image type and the methods listed were tilting, adding noise, rotating, blurring and other ways of modifying the already provided images. This way corrupted images can also be processed by the algorithm and same data distribution is followed while preserving the original labeling.

In [3] authentication based on deep learning model was presented by the authors which used data points from the walking activity data set or the UCI user identification from walking activity data set for making the classification. However, due to the performance improvement that can be achieved by increasing the size of the data set, authors have included the use of 4 different augmentation techniques. Jittering is a technique that adds a mechanical noise that is a part of the sensor readings and in this paper Gaussian noise was applied to the seed data set. Another method that was described is scaling, where the raw data is multiplied by a certain scalar value to change the magnitude of the data set. Rotation was also used as a way of increasing the data set and it is based on the label invariability when sensors are following different orientations. The last technique involves time-warping, a method of stretching or shortening the time series based on its properties.

Another work discusses benefits of data augmentation for continuous authentication, where the additional data can pre-

vent over fitting while covering more input space and leading the model to generalize better during its performance [4]. The data is collected in a matrix form and five data augmentation techniques are employed by the authors including permutation, sampling, scaling, cropping, and jittering. During permutation, the matrix is divided into a certain number of sub matrices and the elements within each sub matrix are shuffled leaving new options from which one modified sub matrix is randomly chosen and added to the original matrix. Sampling the data is performed by introducing a new frequency with which data is picked from the original time curve so the data points in between the readings are picked to create a new variation of the available time series. Scaling is based on the concept of noise being a natural part of sensor readings and introducing a scalar that is between 0.99 and 1.01 can increase the robustness of the data set. Cropping is done by selecting a smaller subset of the data set of certain size with no altering performed on the raw data. Jittering with a noise is also one of the methods the authors used in the paper.

One of the earliest works on time series data augmentation performed for machine learning classification is [5]. The window slicing (WS) approach was used where subsets of the original time series are extracted in a continuous manner where window of a certain length is shifting one data point at a time. Another paper used window warping (WW) in addition to WS [6]. However it is important to point out that during the use of these methods no alterations are made to the value of the data points themselves.

The authors of the survey on time-series data classification [7] point out that there is not enough information on the effect of data augmentation on the performance of the classification algorithms, therefore meaning that the current paper can be a great addition into this research field.

III. Experimental Set Up

The dataset used for the experiment is the UCI user identification subset that belongs to the walking activity dataset. This dataset consists of data collected from twenty-two subjects with averages of total data per subjects ranging from one thousand to five thousand data points. The data is acquired by placing android smartphone in the chest pocket while the subjects are walking on predetermined path. This data set has four features; time-step, x acceleration, y acceleration, z acceleration. By having those features, the dataset represents a human walking pattern. For the current experiment, five subjects and three features are chosen as time-step feature is excluded. Due to variance present in the data per each subject, only 1115 data points are used. The used data is divided into there subsets. There are 555 data points in the training dataset, 280 data points in the test and validation datasets respectively. The amplitude of values varies from negative to positive decimal numbers.

For training part, the WS approach is used for converting training data into training batches. As the window length for each batch is five, the number of obtained batches is the data length of training dataset divided by the window length. A similar batch slicing is used in [8].

IV. Augmentation and Evaluation Methodology

The long short-term memory (LSTM) architecture is selected for the augmentation model. The model is defined to have two layers of bidirectional LSTM. The first layer has six nodes and the second layer has three nodes. Both of them are equipped with RELU as the activation function. For scaling the data, standard scaler is chosen as the dataset contains negative values and it transforms the data to have a standard normal distribution. The optimizer is Adam with 0.001 as the default learning rate and the loss function is mean square error (MSE). The batch size used is five with number of epochs being equal to 3000. The early stopping function is applied as a callback during the training stage, so when the model is optimized or the training and validation loss do not decrease for several epochs the best model is saved and the training stops.

For evaluating the results, the ensemble classifiers are utilized like in [9] where the author uses AdaBoost for classifying the same dataset. In addition, the comparison of standard deviation and mean values extracted from the validation dataset and the resulting augmented data points will be performed to assess the result's validity.

The same model architecture was trained individually for each of the 5 subjects as the data was presented separately for each participant in the original dataset. This resulted in 5 sets of weights that reflected the data distribution of each individual. The other reason for creating a separate model for each subject is a high interclass variation, so combining them for feeding to one model could cause the ability of the model to observe independent patterns for each individual to be corrupt.

V. Result and Augmentation Analysis

Table II shows the training loss, validation loss, accuracy and validation accuracy observed during the training phase. The labels 1 and 5 achieve a good accuracy but in terms of validation accuracy, the label 5 performs worse than the label 2.

Tables III is the comparison of validation data and augmented data in terms of standard deviation and mean values. There are no significant differences between both datasets.

Table I and Figure 1 describe the results of classification using ensemble classifiers. The validation dataset scores have a better average accuracy in all classifiers. However in the Bagged Trees and Subspace KNN the results for the augmented dataset surpass those for the validation dataset.

VI. Conclusion

This paper proposed an augmentation model for biometric dataset that follows LSTM architecture along with the analysis of the resulting performance metrics. As the biometric dataset is hardly accessed by public, the development of M2M biometric - based authentication system had limitations that did not allow for further improvement. Since the biometric data set

TABLE I
CLASSIFICATION ACCURACY AVERAGE OF AUGMENTED DATA SET AND VALIDATION DATA SET

Data Set	Classifier				
	Boasted Trees	**Bagged Trees**	**Subspace Discriminant**	**Subspace KKN**	**RUSBoosted Trees**
Augmented Data	50.91	**99.15**	70.87	**98.95**	50.91
Validation Data	69.1	94.88	90.05	93.76	68.4

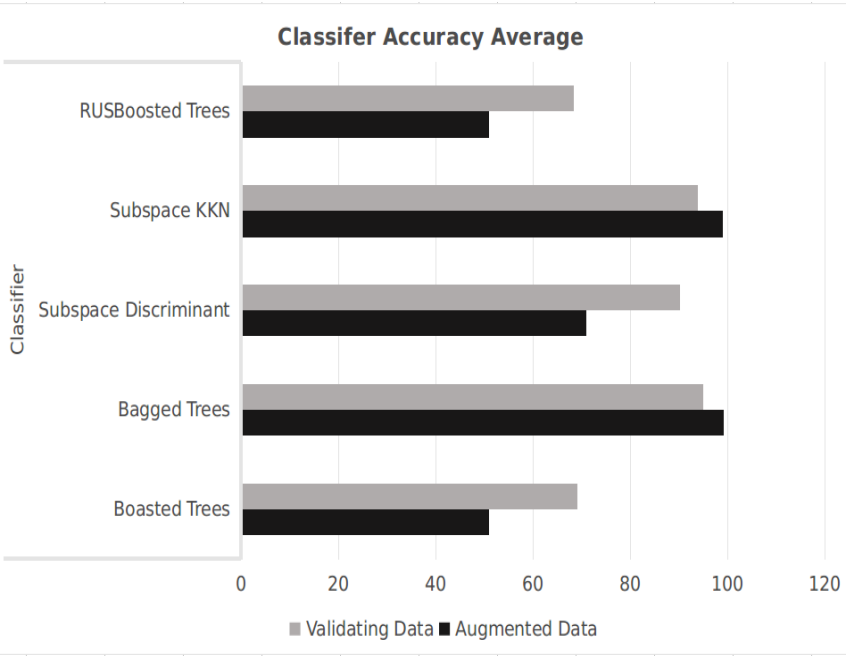

Fig. 1. Classifier Accuracy Average Comparison between Validating Data and Augmented Data

TABLE II
TRAINING LOSS (TL), VALIDATION LOSS (VL), ACCURACY, VALIDATION ACCURACY OF TRAINED AUGMENTATION MODELS FOR FIVE LABELS.

Label	TL	VL	Accuracy	Validation Accuracy
1	**0.0468**	**0.0055**	**0.9438**	**1**
2	0.1937	0.0525	0.8202	0.9130
3	0.1467	0.3443	0.8202	0.6957
4	0.1189	0.4107	0.8427	0.5217
5	**0.0411**	**0.0982**	**0.8764**	**0.7391**

TABLE III
STANDARD DEVIATION AND MEAN COMPARISON BETWEEN VALIDATION DATA AND AUGMENTED RESULT

Label	Validation Data		Augmented Data	
	Standard Deviation	**Mean**	**Standard Deviation**	**Mean**
1	4.178	1.14	5.211	2.084
2	4.982	2.604	5.057	2.794
3	4.024	3.685	4.663	4.173
4	7.649	2.018	6.445	1.708
5	**5.964**	**-1.232**	**3.238**	**3.639**

does not always contain image data, an exclusive evaluation of augmentation techniques is needed for data of different nature. Thus the augmentation model that could solve the problem by employing deep neural network was required. Finally, the augmentation model presented in the paper could mitigate the small dataset problem as is shown by the results.

VII. ACKNOWLEDGMENT

This work was supported by the Institute of Information and Communications Technology Planning and Evaluation (IITP) funded by the Korean Government (MSIT) under Grant 2019-0-00426 (Development of active kill-switch and biomarker-based defence system for life-threatening IoT medical devices).

REFERENCES

[1] R. Nadia, B. A. Tama, and J. Song, "Seamless human impedance-based iot authentication with machine learning techniques," in *2020 International Conference on Information and Communication Technology Convergence (ICTC)*, 2020, pp. 339–343.

[2] P. Corcoran, C. Costacke, V. Varkarakis, and J. Lemley, "Deep learning for consumer devices and services 3—getting more from your datasets with data augmentation," *IEEE Consumer Electronics Magazine*, vol. 9, no. 3, pp. 48–54, 2020.

[3] S. Mekruksavanich and A. Jitpattanakul, "Convolutional neural network and data augmentation for behavioral-based biometric user identification," in *ICT Systems and Sustainability*, M. Tuba, S. Akashe, and A. Joshi, Eds. Singapore: Springer Singapore, 2021, pp. 753–761.

[4] Y. Li, H. Hu, and G. Zhou, "Using data augmentation in continuous authentication on smartphones," *IEEE Internet of Things Journal*, vol. 6, no. 1, pp. 628–640, 2019.

[5] Z. Cui, W. Chen, and Y. Chen, "Multi-scale convolutional neural networks for time series classification," *arXiv preprint arXiv:1603.06995*, 2016.
[6] A. Le Guennec, S. Malinowski, and R. Tavenard, "Data augmentation for time series classification using convolutional neural networks," in *ECML/PKDD workshop on advanced analytics and learning on temporal data*, 2016.
[7] H. I. Fawaz, G. Forestier, J. Weber, L. Idoumghar, and P.-A. Muller, "Deep learning for time series classification: a review," *Data Mining and Knowledge Discovery*, vol. 33, no. 4, pp. 917–963, 2019.
[8] C. Oh, S. Han, and J. Jeong, "Time-series data augmentation based on interpolation," *Procedia Computer Science*, vol. 175, pp. 64–71, 2020, the 17th International Conference on Mobile Systems and Pervasive Computing (MobiSPC),The 15th International Conference on Future Networks and Communications (FNC),The 10th International Conference on Sustainable Energy Information Technology. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1877050920316914
[9] P. Casale, O. Pujol, and P. Radeva, "Personalization and user verification in wearable systems using biometric walking patterns," *Personal and Ubiquitous Computing*, vol. 16, no. 5, pp. 563–580, Jun 2012. [Online]. Available: https://doi.org/10.1007/s00779-011-0415-z