Original Data Vs High Performance Augmented Data for ANN Prediction of Glycemic Status in Diabetes Patients

Massaro, Alessandro and Magaletti, Nicola and Giardinelli, Vito O. M. and Cosoli, Gabriele and Leogrande, Angelo and Cannone, Francesco

Lum University-Giuseppe Degennaro; Lum Enterprise s.r.l., Lum University-Giuseppe Degennaro; Lum Enterprise s.r.l., Lum University-Giuseppe Degennaro; Lum Enterprise s.r.l., Lum University-Giuseppe Degennaro; Lum Enterprise s.r.l., Lum University-Giuseppe Degennaro; Lum Enterprise s.r.l.

5 April 2022

Alessandro Massaro[1], Nicola Magaletti[2], Vito O.M. Giardinelli[3], Gabriele Cosoli[4], Angelo Leogrande[5], Francesco Cannone[6]

Original Data Vs High Performance Augmented Data for ANN Prediction of Glycemic Status in Diabetes Patients

Abstract

In the following article a comparative analysis between Original Data (OD) and Augmented Data (AD) are carried out for the prediction of glycemic status in patients with diabetes. Specifically, the OD concerning the time series of the glycemic status of a patient are compared with AD. The AD are obtained by the randomised average with five different ranges, and are processed by a Machine Learning (ML) algorithm for prediction. The adopted ML algorithm is the Artificial Neural Network (ANN) Multilayer Perceptron (MLP). In order to optimise the prediction two different data partitioning scenarios selecting training datasets are analysed. The results show that the algorithm performances related to the use of AD through the randomisation of data in different ranges around the average value, are better than the OD data processing about the minimization of statistical errors in self learning models. The best achieved error decrease is of 75.4% if compared with ANN-MLP processing of the original dataset. Furthermore, in the paper is added a linked discussion about the economic and managerial impact of AD in the healthcare sector.

Keyword: *ANN-Artificial Neural Network, Augmented Data Generation, Telemedicine, E-Healthcare, Model Optimization.*

1.Introduction-Research Question

The proposed approach considers the comparison of Artificial Neural Network (ANN) Multilayer Perceptron (MLP) algorithm performance using the basic dataset, named Original Data (OD), and Augmented Data (AD) generated by a method based on a randomization process. Specifically, the ANN algorithm is applied for the prediction of the glycemic status of patients with diabetes. The proposed approach is based on the randomisation data process around the averag++e value of the OD. The randomisation process is estimated in a range expressed as the percentage variation from the average value. The different percentages are useful to define the best AD dataset according to original

[1]Professor at Lum University Giuseppe Degennaro, and Chief Research Officer-CRO at Lum Enterprise s.r.l. Email: massaro@lum.it. Strada Statale 100 km 18, 70010 Casamassima BA, Puglia, Italy, European Union.

[2] Chief Operation Officer-COO and Senior Researcher at Lum Enterprise s.r.l. Email: magaletti@lumenterprise.it. Strada Statale 100 km 18, 70010 Casamassima BA, Puglia, Italy, European Union.

[3] Business Developer and Researcher at Lum Enterprise s.r.l. Email: giardinelli@lumenterprise.it Strada Statale 100 km 18, 70010 Casamassima BA, Puglia, Italy, European Union.

[4] Senior IT Specialist and Solutions Architects and Researcher at LUM Enterprise s.r.l. Email: cosoli@lumenterprise.it. Strada Statale 100 km 18, 70010 Casamassima BA, Puglia, Italy, European Union.

[5]Assistant Professor at Lum University Giuseppe Degennaro and Researcher at Lum Enterprise s.r.l. Email: leogrande.cultore@lum.it. Strada Statale 100 km 18, 70010 Casamassima BA, Puglia, Italy, European Union.

[6] Emtesys s.r.l, Piazza Giuseppe Massari, 6 – 70122 Bari (BA), Italy; info@emtesys.com

dataset oscillations. The use of AD is relevant as it allows to make predictions maximising the information capacity of the ANN MLP algorithm and, consecutively, improving the self-learning model. In fact, in cases in which it is necessary to create predictive scenarios having few data, the use of AD can optimise the data processing results by the means of the optimization of the training dataset model. Using AD automatically generated with specific tools such as Node-RED, it is possible to automate the data processing thus providing a usable data driven solution suitable for telemedicine information systems. The comparison between ANN-MLP performance with OD and OD plus AD is performed by considering the average of the historical series of the original randomised data in a range between a specific percentage. Furthermore, are used two different hyperparameters that are "*80-20 Take from The Top"* (80-20 FTT), where the testing data are AD listed in top of the database representing the last estimated data, and "*80-20 Linear Sampling*" (80-20 LS), where selecting the testing rows linearly over the whole table e.g. every third row. The results confirm that using AD, it is possible to obtain a significant minimization of the mean of statistical errors that in the case of 80-20 FTT is equal to -75.4% in respect to original data. The results show that the ANN can be optimised using AD. It therefore follows that the AD model performs even better than the original time series in terms of prediction performance. This increase makes it possible to extend the use of ML and predictive models even regardless of the OD. The result is therefore an increase in the ability to build scenarios that can be useful to prevent health critical situations also when data are insufficient or incomplete.

2. Literature Review

An analysis of the scientific literature relating to the use of AD is presented below.

2.1 AD methodology for various applications

[1] employ the AD to evaluate the commitment of university teachers. [2] applies AD for intelligent river maintenance. [3] generate AD using digital twins for customization. [4] refer to the use of AD to compensate for analytical situations characterised by the absence or insufficiency of data. The analysis shows that by means of AD it is possible to create a Support Vector Machine (SVM) classifier that performs better than Recursive Neural Network (RNN) and Deep Neural Network (DNN) algorithms. Authors in [5] apply the AD for the prediction of risks associated with autonomous driving. Researchers in [6] introduce the concept of Counterfactual Augmented Data (CAD) to realise a model within the framework of the Natural Language Processing (NLP). The authors verified that the use of AD has increased the ability to discard words with reduced semantic value and to act better out of domain. [7] present a case of application of AD to the insurance sector. It is also possible to use AD to predict sales in the retail sector [8], and to evaluate the socio-economic determinants of human capital experts in Information Technology (IT) disciplines [9]. Authors in [10] use CAD in the context of text mining analysis to distinguish between main features and artefact data. [11] use a three-operation model to generate AD i.e. flipping, translation and rotation. [12] use AD to predict depression from social media posting. They refer to the use of AD for deep learning with attention to the issue of data quality and compliance with fidelity, variety, and veracity.

2.2 AD in healthcare and telemedicine

There are many applications of augmented data in healthcare [13], [14], [15], [16], [17]. [18] apply AD to solve the question of missing data in healthcare with a Bayesian approach. [19] use AD to analyse heartbeats by training a convolutional network and achieving a significant level of accuracy. AD was also tested [20] in the analysis of colorectal histopathological images. [21] employ of AD is mostly utilised in deep learning applications to medical image analysis. In fact, in this case the possibility of training algorithms meets the limit of data availability. Applying AD in association with

deep learning techniques, it is possible to obtain significant results in predictive and accuracy terms. [22] apply AD to the optimization of health services offered through smartphones. [23] propose an analytical model aimed at the application of the Multiple Additive Regression Trees (MART) algorithm for the classification of diseases with the use of AD for the reduction of class imbalances. The application of the AD leads to a significant increase in the performance of the MART. [24] apply AD to predict the level of depression through the usage of health applications [25] implement AD for the analysis of Electronic Health Records (EHR) created using convolutional neural networks (CNN). [26] adopt the AD to carry out a dataset completion activity in the case of rare diseases. In fact, rare diseases give rise to insufficient data. Therefore, the data contained in the electronic medical records relating to rare diseases are increased in order to carry out predictive analyses applying ML algorithms. [27] propose AD to make up for the lack of data in medical records and to subsequently be able to apply Reinforcement Learning (RL) algorithms. [28] apply AD to diabetes prediction. [29] generate AD to complement real data for Covid-19 case prediction. Finally, in [30] are applied AD for breast cancer prediction.

2.3 Methodology

The software used for the prediction is Konstanz Information Miner (KNIME) [31]. The activity of randomization of data has been realised using the Node-RED [32], [33] tool able to automate AD generation. Node-RED is a tool able to connect hardware, with software and online services in the perspective of the Internet of Things (IoT). Node-RED therefore allows you to generate connections centred on the exchange of information between physical devices and software. Node-RED is software that was introduced by IBM. The Node-RED editor simulates the operation of a web browser. The choice of using Node-RED is due to its simplicity of data modifications that allow it to generate AD automatically. Node-RED allows modifying the input data files by applying formulas as for the randomisation data processing of the proposed approach.

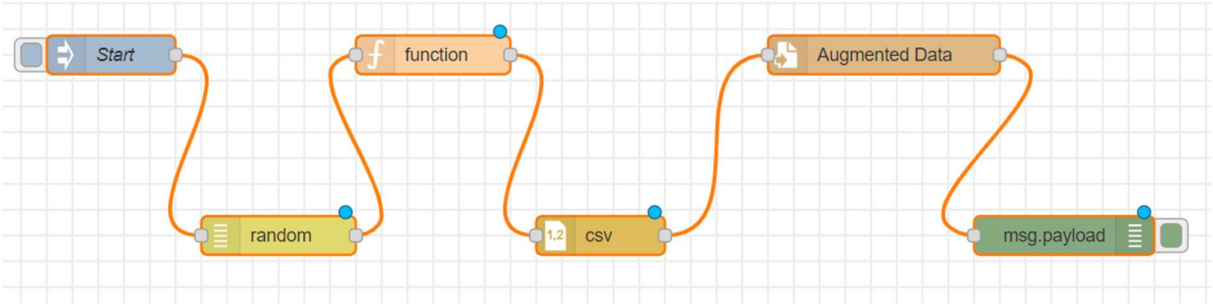

Figura 1. Example of the Node-RED workflow generating AD.

The flow described in the figure is made up of six nodes:

- *Start*: starting node in which to set the time range for each detection-in this case 5 seconds;
- *Random*: setting of randomization parameters;
- *Function*: setting script to define the percentage thresholds;
- *CSV*: creation of the output csv containing the simulation result;
- *Augmented Data***:** creation of the structure of the final output that allows the correct reading of the csv.

The flow was used to automate the dataset generation procedures. The whole dataset, composed of OD and AD, is processed as an excel input file. The KNIME workflow is able to process by ANN-MLP the whole dataset. To obtain the augmented data, different methods were used, namely:

- *the average of the historical series of the original randomised data in a range between -3% and + 3%;*
- *the average of the historical series of the original randomised data in a range between -4% and + 4%;*
- *the average of the historical series of the original randomised data in a range between -5% and + 5%;*
- *the average of the historical series of the original randomised data in a range between -6% and + 6%;*
- *the average of the historical series of the original randomised data in a range between -7% and + 7%.*

The different percentages are chosen to find the best ANN-MLP performance according to the original dataset oscillation behaviour. The record number of the training dataset is increased by AD. An approach to consider is the data randomization around the average value of the few available data-basic dataset [34]. The randomization can be performed between thresholds of about ± *x*%, where *x* can be optimised in function of the oscillation behaviour of the original dataset. In Fig. 2 is sketched the proposed approach used for the comparison.

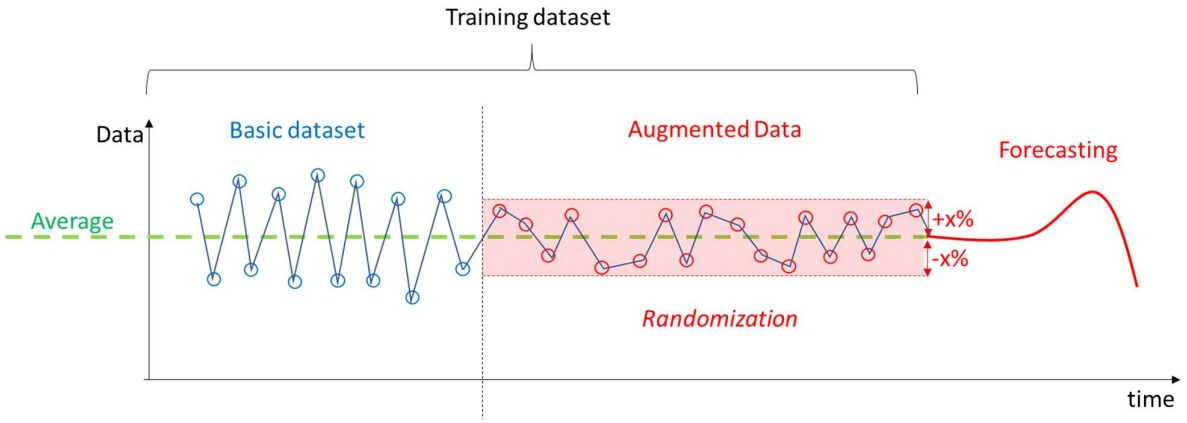

Figure 2. Example of AD generated from a basic dataset.

The whole adopte methodology is illustrated in Fig. 3 where it is possible to distinguish the following main functions:

- *Glucometer*: device saving digital data of glucose measurement;
- *OD dataset:* the original dataset is stored into a cloud database;
- *data pre-processing*: some outliers are cleaned to optimise data quality for AD creation;
- *Node-Red workflow*: generating AD by means the randomisation approach setting threshold percentage;
- *KNIME workflow* executing ANN-MLP algorithm-the whole dataset is partitioned into a training and a testing dataset-and reading as input a constructed excel file merging AD of the Node-Red workflow with OD, see Fig. 4;
- *Estimation of performance indicators*: to set the best percentage threshold for the AD generation- see Appendix A;

- *Glucose prediction:* glucose parameter forecasting as output of the KNIME ANN-MLP algorithm.

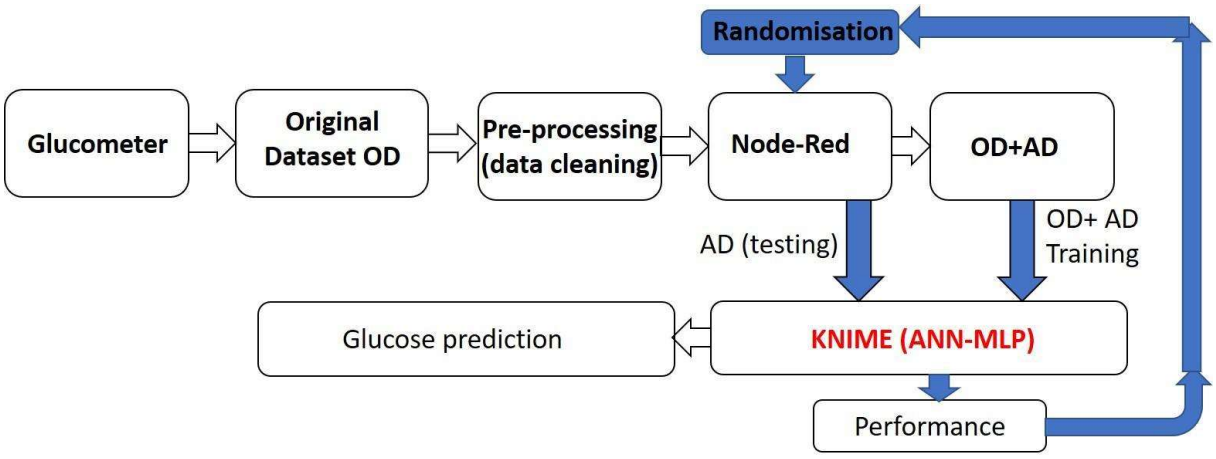

Figure 3. Full applied Methodology used to check ANN-MLP performance.

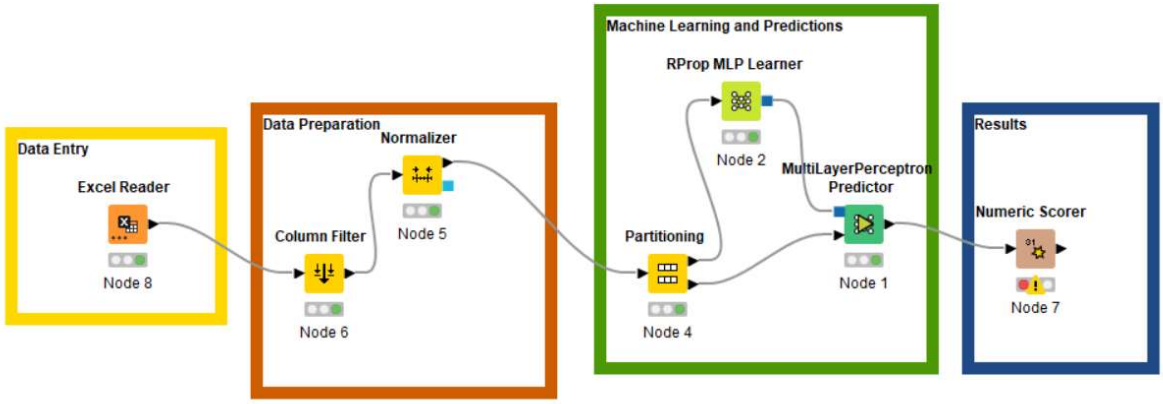

Figure 4. Methodology applied to check ANN-MLP performance.

3. ANN-MLP experimental results

Below are discussed all the performed experimental tests summarised in the appendix A.

3.1 Hyperparameters Take From The Top

The values relating to the randomization of the data around the mean through the selection of the 80-20 with the hyper parameter Take from the Top are reported below. That is, 80% of the data was used for learning the algorithm and the remaining 20% was used for the actual prediction.

3.1.1 Randomization around the average of ± 3%

The values of the statistical errors relating to randomization are shown below with an average of plus or minus 3%:

- *Decrease in the Mean Absolute Error equal to -52.076%;*
- *Reduction of the Mean Squared Error equal to -100.00%;*
- *Reduction of the Root Mean Squared Error with a variation of -56.447%;*

The average statistical errors obtained through the randomization of the mean value around plus or minus 3% was equal to a value of -69,50%. There was therefore a significant decrease in statistical errors.

3.1.2 Randomization around the average of ± 4%

The data relating to the predictions made with randomization around an average of the statistical error of between plus and minus 4% are reported below.

- *Reduction of the Mean Absolute Error by an amount equal to -50.448%;*
- *Reduction of the Mean Squared Error equal to an amount of -100.00%;*
- *Reduction of the Root Mean Squared Error equal to a value of -54.56%;*

The mean values of the statistical errors of randomization around the mean value of plus or minus 4% is equal to a value of -68,33%.

3.1.3 Randomization around the average of ± 5%

The value of randomised statistical errors around the mean value with plus or minus 5%:

- *Decrease in the Mean Absolute Error with a variation equal to a value of -5.27%;*
- *Decrease in the Mean Absolute Error equal to an amount of -36.5%;*
- *Reduction of the Mean Squared Error equal to a value of -20.338%;*

On average, the value of statistical errors in the case of randomization around 4% is equal to a value of -20.717%.

3.1.4 Randomization around the average of ± 6%

The following are the data relating to the prediction made with the randomization around a value of plus or minus 6%:

- *Decrease in the value of the Mean Absolute Error with a reduction equal to an amount of-71.859%;*
- *Reduction of the Mean Squared Error with a variation equal to a value of -100.00%;*
- *Root Mean Squared Error with a decrease equal to an amount of -54.462%;*

The average value of statistical errors decreased by an amount equal to -75.440%.

3.1.5 Randomization around the average of ± 7%

The statistical errors relating to the prediction with the randomization of the mean with values varying between plus or minus 7% are indicated below. The results are shown below i.e .:

- *Decrease in the Mean Absolute Error with a change equal to an amount of -40.904%;*
- *Reduction of the Mean Squared Error with a variation equal to an amount of -48.2%;*
- *Reduction of the Root Mean Squared Error with a change equal to an amount of -45.355%;.*

Overall, the average value of the statistical errors is equal to -44.80%.

3.2 80-20 Linear Sampling

The values relating to the randomization of the data around the mean through the selection of the 80-20 Linear Sampling are reported below. That is, 80% of the data was used for learning the algorithm and the remaining 20% was used for the actual prediction.

3.2.1 Randomization around the average of ± 3%

The value of the statistical errors in the case of randomization with an average value of around plus or minus three% is analysed below:

- *Decrease in the Mean Absolute Error with a variation of -39.537%;*
- *Reduction of the Mean Squared Error with a variation equal to a value of -46.420%;*
- *Reduction of the Root Mean Squared equal to a value of -26.802%;*

On average, the value of statistical errors decreased by a value equal to -37.585%.

3.2.2 Randomization around the average of ± 4%

Below is the value of the statistical errors of the predictions with the average value around 4% compared to the original value:

- *Decrease in the Mean Absolute Error equal to a change equal to a change of -38.198%;*
- *Reduction of the Mean Squared Error equal to a variation equal to a variation equal to a value of -45.365%;*
- *Reduction of the Root Mean Squared Error equal to a variation equal to a value of -26.085%;*

On average, the value of statistical errors decreased by an amount equal to a value of -36.549% compared to the original value.

3.2.3 Randomization around the average of ± 5%

The statistical errors in relation to the prediction made on the randomised series with an amount equal to plus or minus 5% compared to the original value are analysed below, that is:

- *Reduction of the Mean Absolute Error equal to an amount of -33.688%;*
- *Reduction of the Mean Squared Error equal to a variation of -41.258%;*
- *Reduction of the Root Mean Squared Error equal to an amount of -23.357%;*

On average, the value of the statistical errors calculated with the randomised prediction around an average ±5% was equal to a value of -32.7674%.

3.2.4 Randomization around the average of ± 6%

The statistical errors made by using the randomised series around the value of plus or minus 6% are analysed below;

- *Reduction of the Mean Absolute Error equal to an amount of -35.591%;*
- *Reduction of the Mean Squared Error equal to an amount of -41.396%;*
- *Reduction of the Root Mean Squared Error equal to an amount of -23.447%;*

On average, the value of the statistical errors decreased by an amount equal to a value of -33.477 in the use of the randomised series with variations on the average of between plus or minus 6% compared to the value of the original series.

3.2.5 Randomization around the average of ± 7%

The value of the percentage variation of the statistical errors obtained using randomization around the value of plus or minus 7% compared to the original values is analysed below, that is:

- *Decrease in the value of the Mean Absolute Error by an amount equal to -36.366%;*
- *Reduction of the Mean Squared Error by an amount equal to a value of -44.55%;*

- *Reduction of the Root Mean Squared Error with a variation equal to an amount of -25.539%;*

On average, the value of the statistical errors relating to the prediction with the randomised series around the average value with variations of plus or minus 7% was equal to an amount of -35.486%.

4. Strategy to Optimise the Prediction Ability of ANN-MLP

The techniques that have been applied to the processing of the OD, to the data augmentation process and to the efficiency of the hyperparameters for the optimization of the predictive capacity achieved using the ANN-MLP algorithm are briefly analysed below. In fact, the performances of the ANN-MLP algorithm are highly variable based on the hyperparameters used. The activity of hyperparameters optimization therefore becomes essential to make the most of the predictive capabilities of the ANN-MLP algorithm especially if used for the evaluation of similar datasets with marginal differences such as those used in the case presented. In this sense, it must be considered that the use of KNIME software offers a set of nodes and tools to achieve hyperparameters optimization that allow to significantly increase the efficiency of the prediction in minimising the main statistical errors.

- *Description of the dataset:* the dataset used refers to measurements of the glycemic status of a patient with diabetes. Specifically, the original dataset used is made up of about 13,000 data, with respect to which data was added using AD. Specifically, the measurements that have been acquired have a temporal distance among them of approximately 3 minutes. Therefore, the dataset represents the glycemic state of a patient with diabetes with a survey carried out every 3 minutes. The dataset used has about 13,099 records. The average of the surveys is equal to 131.22 units, the median is equal to 118.00 units, the minimum value is equal to 14.00 units, the maximum value is equal to 491.00 units. The standard deviation is equal to 49.785, the asymmetry value is equal to 1.353 and the kurtosis value is equal to an amount of 3.08. These data were subsequently processed, purified of the absolute minimum and maximum elements and subjected to the method of creating increased data using the distribution mean
- *Data preparation:* before using records for our analysis, we have eliminated outliers' values from our dataset generated by potential anomalies in the data-recording phase. We cleaned our dataset from values lower than 10 and higher than 500 in the sense of glycemic status. Anyway, outliers were only 1% of records. The elimination of the outliers is necessary to give the dataset a structure that can be subjected to the increase of data. In fact, in the event of the creation of increased data using the entire data dataset without eliminating the outliers, the entire data augmentation activity would have been incorrect, as it would have had in its distribution also valueswith a low probability. It follows therefore that the realisation of an AD model is more credible in the presence of a set of data that, purified of the outliers, actually represents the data that can be assumed in the analysed case study.
- *Partitioning:* The use of the hyperparameters of the partitioning node was an essential element of the entire research strategy aimed at selecting the best augmented data models in terms of performance in minimising essential statistical errors. In particular, two different partitioning configurations were used. The partitioning phase, within KNIME, is essential as in this node it is possible to determine the level of learning rate, i.e. the percentage of data that is used to train the algorithm. Specifically, in the case analysed, the value of the learning rate was set at an amount of 80% of the inserted dataset. As a result, 20% of the data in the dataset was used for prediction. After identifying the percentage of the learning rate, two different partitioning configurations were used alternatively: FTT and LS. Using the FTT mode, the topmost data

is sent to the learner while the remaining data are used for the actual prediction. The application of the LS mode is carried out in a sampling activity in which the first and last row are always included in the dataset while the intermediate values are organised in a linear way.

Using the techniques indicated it was therefore possible to make the best use of the predictive and performance possibilities of the ANN-MLP algorithm. It must be considered that all the options available in the KNIME nodes have not been used, neither with reference to the use of partitioning nor in the dimension of deep learning with an increase in layers and hidden layers per neuron. In fact, in the approach that was used, we wanted to give greater importance to the methodology of producing the AD through the techniques of data preparation and randomization of the values around the mean. In fact, the focus of the research question was completely aimed at the creation of an augmented dataset that could be used for prediction. In the research strategy used, it was therefore chosen to give greater importance to the process of creating AD rather than to prediction methodologies with ML algorithms. Obviously, since deep learning is a more performing predictive technique in the simple ANN network, it follows that '*a fortiori'* the results obtained with the ANN moving network with the increased data are better performing in a deep learning context.

5. The Economic and Managerial Impact of AD in Healthcare

The use of augmented data in telemedicine also has a significant impact from an economic point of view and from a managerial point of view.

- *Economic effects of the use of AD***.** The economic repercussions of the use of augmented data in telemedicine mainly concern savings for health administrations. In fact, by making predictions using augmented data, it is possible to calculate, for example, the probability of the manifestation of pathologies that can have a significant impact in terms of public spending, as for example happens for silent killers, that is: hypertension and diabetes. In fact, they can somehow anticipate the manifestation of a pathology by creating the conditions for a broad prevention activity on patients. In fact, using AD, the few data that may be available using devices can be adequately used to obtain drinking and own predictions that can help citizens to have higher levels of living conditions with a positive impact also in terms of reducing health care costs**.**
- *Managerial effects of the use of augmented data in telemedicine.* AD have a very significant impact also in managerial terms. AD in the telemedicine sector can be used both in the Business to Business (B2B) sector and by healthcare professionals as well as by healthcare institutions. The value of AD in telemedicine can be very useful for maximising the efficiency of public institutions and companies operating in the health sector. The greatest advantages can therefore be offered in the sense of increasing the production efficiency of health companies, public health institutions and even health professionals. The AD can therefore be used both as diagnostic tools and also for monitoring activities relating to the condition of patients. Furthermore, it is possible to increase the forecasting capacity in a broad sense by highlighting the pathologies that patients could encounter. Finally, with predictions, it is possible to act on changes in patient behaviour to prevent any predicted pathologies using AD.

It follows that the use of augmented data can have important impacts in economic terms and also in managerial terms for healthcare companies, healthcare professionals and healthcare institutions.

7. Hyperparameter ranking for choosing the best configuration

A comparison activity is carried out below in terms of performance between the various configurations of the hyperparameters. The hyperparameters are of two types:

- *related to partitioning techniques:* two different types of partitioning were used, namely 80-20 Linear Sampling and 80-20 From the Top;
- *related to the output of randomization carried out around the value of the mean of the original data-OD*: in this sense, augmented data were created with a value of + or -3%, + or - 4% respectively; + or - 5%, + or -6%, + or - 7%.

$$Ranking_{AD,Hs} = \frac{\left[\left(\frac{MAE_{AD,H_S}}{MAE_{OD}}\right)*100-100\right]+\left[\left(\frac{MSE_{AD,H_S}}{MSE_{OD}}\right)*100-1\right]+\left[\left(\frac{RMSE_{AD,H_S}}{RMSE_{OD}}\right)*100-1\right]}{3} \quad (1)$$

$$Where\ AD = Augmented\ Data, \quad Hs = Hyperparameters\ , \quad OD = Original\ Data$$

Both the FTT-From the Top and LS-Linear Sampling values are reported below. The data are organised in terms of the percentage difference between the value found between the OD and the values found in the AD. Below is the ranking of the configurations of the augmented data with indications of the hyperparameters:

- *Plus or minus 6% 80-20 FTT* -75.4406483;
- *Plus or minus 3% 80-20 FTT* -69.5077721;
- *Plus or minus 4% 80-20 FTT* -68.3362558;
- *Plus or minus 7% 80-20 FTT* -44,8063051;
- *Plus or minus 3% 80-20 LS* -37,5859102;
- *Plus or minus 4% 80-20 LS* -36.5492287;
- *Plus or minus 7% 80-20 LS* -35,4864969;
- *Plus or minus 6% 80-20 LS* -33.4777469;
- *Plus or minus 5% 80-20 LS* -32.7674552;
- *Plus or minus 5% 80-20 FTT* -20.7173404.

Ranking of the Hyperparameters with the Difference between the Original Data-OD and Augmented Data-AD		
Rank	**Hyperparameters**	**Reduction of statistical errors through the use of augmented data**
1	*Plus or minus 6 80-20 FTT*	-75,4406843
2	*Plus or minus 3 80-20 FTT*	-69,5077721
3	*Plus or minus 4 80-20 FTT*	-68,3362558
4	*Plus or minus 7 80-20 FTT*	-44,8063051
5	*Plus or minus 3 80-20 LS*	-37,5859102
6	*Plus or minus 4 80-20 LS*	-36,5492287
7	*Plus or minus 7 80-20 LS*	-35,4864969
8	*Plus or minus 6 80-20 LS*	-33,4777469
9	*Plus or minus 5 80-20 LS*	-32,7674552
10	*Plus or minus 5 80-20 FTT*	-20,7173404

Figure 5. Ranking of hyperparameters for the ability to reduce statistical errors through the use of augmented data.

The predictions that have been made through the use of the most efficient algorithm configurations are shown below. In the following image, the prediction is related to the value of the original historical data through the use of the 80-20 Take From the Top-TFF hyperparameters in the range plus or minus 6%. As is evident from the analysis carried out, the prediction is represented by a linear trend.

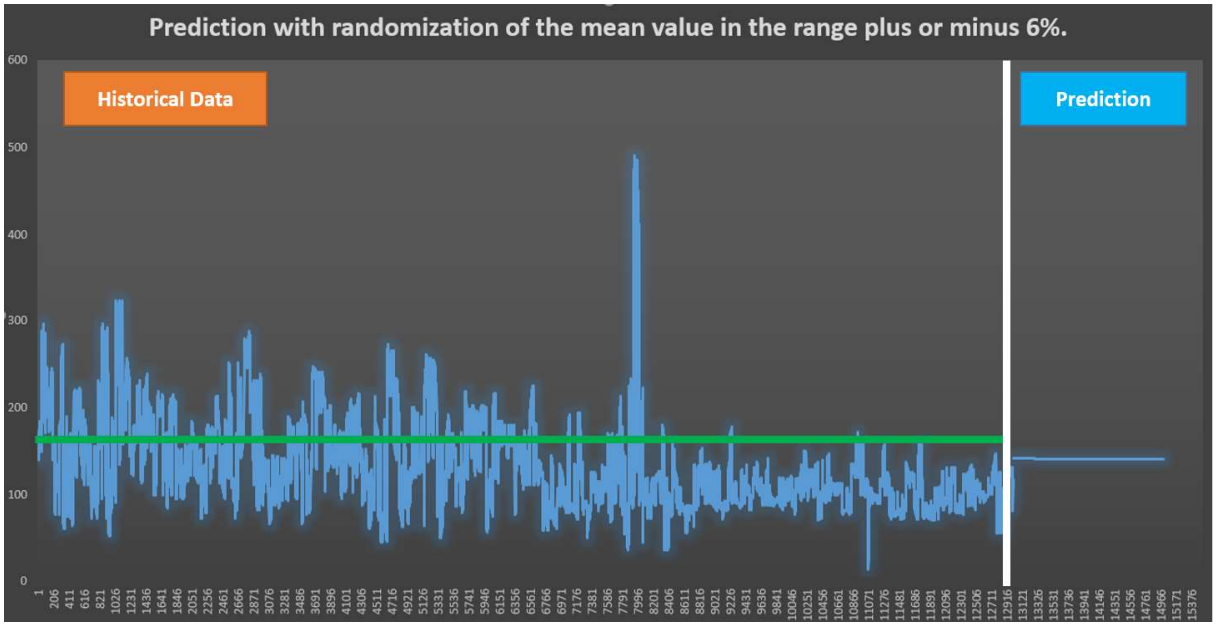

Figure 6. The relationship between historical data and prediction with the partitioning at 80-20 Take from the Top-TFF.

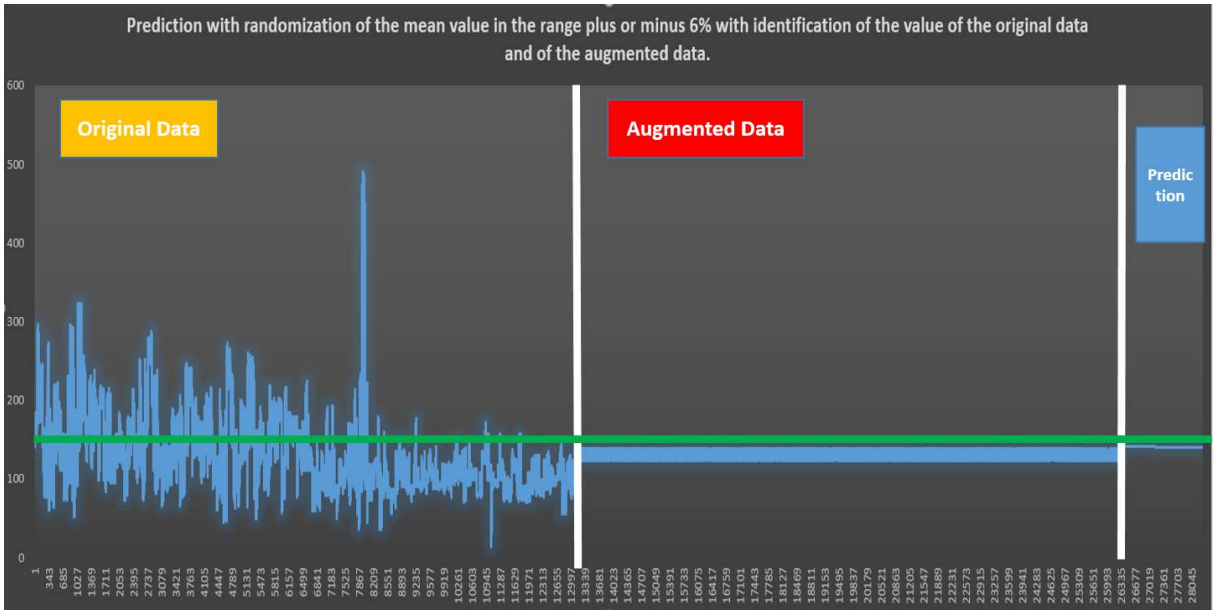

Figure 7. Prediction in connection with original data and augmented data with the hyperparameters 80-20 Take from the Top in the configuration of the randomization around the value of plus or minus 6%.

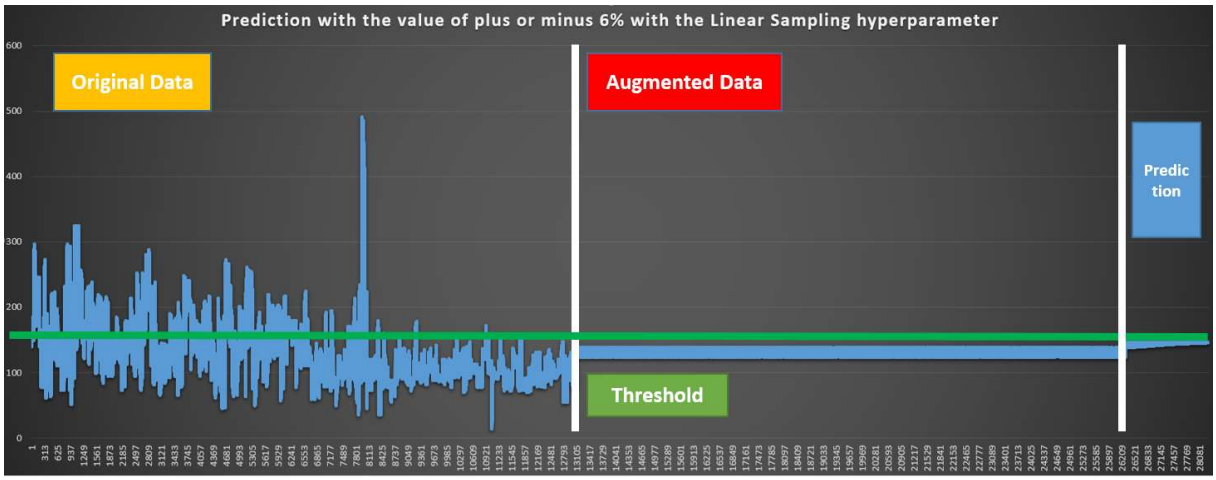

Figure 8. Relationship among original data, augmented data and prediction with the hyperparameters 80-20 Linear Sampling with the methodology of the randomization around the average of plus or minus 6%.

As it is possible to verify by using the data of the prediction, it results that the predicted value is substantially an approximation of a straight line or rather tends to be represented by a vector of a constant. It follows that the prediction tends to be an approximation of a number or of marginal variations in the neighbourhood of a certain number.

7. Conclusions

In summary, this article analyses the predictive performance of algorithms analysed using AD predicting the value of the glycemic status of a patient with diabetes. The patient's original time series data was compared with the AD. Node-Red software was used to generate AD. Five different AD dataset are produced by randomising the average value of the AD with the addition of a range. The

used ranges are: plus or minus 3%, plus or minus 4%, plus or minus 5%, plus or minus 6%, plus or minus 7%. The AD are merged with the OD generating a high performance training dataset model useful for the glycemic prediction by ANN-MLP algorithm. In order to make a comparison between various methodologies, two different data partitioning approaches are used. Analyses show that the AD performs better than the OD in terms of minimising statistical errors. By averaging between a set of statistical errors that is Mean Absolute Error, Means Squared Error, Root Mean Squared Error, Mean Absolute Percentage Error it turns out that predictions with AD are more efficient than predictions with OD. Using the hyperparameters we find that the best results in terms of minimization of statistical errors are associated with 80-20 Take From the Top with a randomization around the average of plus or minus 6 with a value of -75.44%. Finally, the paper discusses the economic and managerial impact of AD in the healthcare scenario.

Figure Index

Bibliography

[1] M. Perkmann, R. Fini, J. M. Ross, A. Salter, C. Silvestri e V. Tartari, «Accounting for universities' impact: Using augmented data to measure academic engagement and commercialization by academic scientists,» Research Evaluation, vol. 24, n. 4, pp. 380-39, 2015.

[2] R. Pierdicca, E. Frontoni, P. Zingaretti, A. Mancini, E. S. Malinverni, A. N. Tassetti e A. Galli, «Smart maintenance of riverbanks using a standard data layer and Augmented Reality,» Computers & Geosciences, vol. 95, pp. 67-74, 2016.

[3] X. Wang, Y. Wang, F. Tao e A. Liu, «New paradigm of data-driven smart customisation through digital twin,» Journal of manufacturing systems, vol. 58, pp. 270-280, 2021.

[4] L. Yi e M. W. Mak, «Adversarial data augmentation network for speech emotion recognition,» 2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), n. IEEE, pp. 529-534, 2019.

[5] S. Y. Yu, A. V. Malawade, D. Muthirayan, P. P. Khargonekar e M. A. Al Faruque, «Scene-graph augmented data-driven risk assessment of autonomous vehicle decisions,» IEEE Transactions on Intelligent Transportation Systems, 2021.

[6] D. Kaushik, A. Setlur, E. Hovy e Z. C. Lipton, «Explaining the efficacy of counterfactually augmented data,» arXiv preprint arXiv:2010.02114., 2020.

[7] A. Massaro, A. Panarese, M. Gargaro, A. Colonna e A. Galiano, «A Case Study of Innovation in the Implementation of a DSS System for Intelligent Insurance Hub Services,» Comput. Sci. Inform. Technol, vol. 9, pp. 14-23, 2021.

[8] A. Massaro, A. Panarese, D. Giannone e A. Galiano, «Augmented Data and XGBoost Improvement for Sales Forecasting in the Large-Scale Retail Sector,» Applied Sciences, vol. 7793, n. 17, p. 11, 2021.

[9] A. Leogrande, N. Magaletti, G. Cosoli, V. Giardinelli e A. Massaro, «ICT Specialists in Europe.,» University Library of Munich Germany, 2022.

[10] I. Sen, M. Samory, F. Floeck, C. Wagner e I. Augenstein, «How Does Counterfactually Augmented Data Impact Models for Social Computing Constructs?,» arXiv preprint arXiv:2109.07022, 2021.

[11] S. Dargan, M. Kumar, M. R. Ayyagari e G. Kumar, «A survey of deep learning and its applications: a new paradigm to machine learning,» Archives of Computational Methods in Engineering, vol. 27, n. 4, pp. 1071-1092, 2020.

[12] S. Kayalvizhi e D. Thenmozhi, «Data set creation and empirical analysis for detecting signs of depression from social media postings,» arXiv preprint arXiv:2202.03047, 2022.

[13] A. Massaro, «Information Technology Infrastructures Supporting Industry 5.0 Facilities,» Electronics in Advanced Research Industries: Industry 4.0 to Industry 5.0 Advances, vol. IEEE, n. 10.1002/9781119716907.ch2., pp. 51-101, 2022.

[14] A. Massaro, V. Maritati, N. Savino e A. Galiano, «Neural Networks for Automated Smart Health Platforms oriented on Heart Predictive Diagnostic Big Data Systems,» AEIT International Annual Conference, n. 10.23919/AEIT.2018.8577362., pp. 1-5, 2018.

[15] A. Massaro, A. Galiano, D. Scarafile, A. Vacca, A. Frassanito, A. Melaccio e F. Attivissimo, «Telemedicine DSS-AI multi level platform for monoclonal gammopathy assistance,» IEEE International Symposium on Medical Measurements and Applications (MeMeA), pp. 1-5, 2020.

[16] A. Massaro, G. Ricci, S. Selicato, S. Raminelli e A. Galiano, «Decisional Support System with Artificial Intelligence oriented on Health Prediction using a Wearable Device and Big Data,» IEEE International Workshop on Metrology for Industry 4.0 & IoT, 2020.

[17] A. Massaro, V. Maritati, N. Savino, A. Galiano, D. Convertini, E. De Fonte e M. Di Muro, «A study of a health resources management platform integrating neural networks and DSS telemedicine for homecare assistance,» Information, vol. 7, n. 176, p. 9, 2018.

[18] S. Kouchaki, N. Pourshahrokhi, K. M. Kober, C. Miaskowski e P. Barnaghi, «A Hybrid Bayesian Model to Analyse Healthcare Data,» 35th Conference on Neural Information Processing Systems , 2021.

[19] U. R. Acharya, S. L. Oh, Y. Hagiwara, J. H. Tan, M. Adam, A. Gertych e R. San Tan, «A deep convolutional neural network model to classify heartbeats,» Computers in biology and medicine, vol. 89, pp. 389-396, 2017.

[20] J. Wei, A. Suriawinata, L. Vaickus, B. Ren, X. Liu, J. Wei e S. Hassanpour, «Generative image translation for data augmentation in colorectal histopathology images,» Proceedings of machine learning research, vol. 10, p. 116, 2019.

[21] P. Chlap, H. Min, N. Vandenberg, J. Dowling, L. Holloway e A. Haworth, «A review of medical image data augmentation techniques for deep learning applications,» Journal of Medical Imaging and Radiation Oncology, vol. 65, n. 5, pp. 545-563, 2021.

[22] B. Longstaff, S. Reddy e D. Estrin, «Improving activity classification for health applications on mobile devices using active and semi-supervised learning,» 2010 4th International Conference on Pervasive Computing Technologies for Healthcare, vol. IEEE, pp. 1-7, 2010.

[23] V. C. Pezoulas, T. P. Exarchos, A. G. Tzioufas e D. I. Fotiadis, «Multiple additive regression trees with hybrid loss for classification tasks across heterogeneous clinical data in distributed environments: a case study,» 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), vol. IEEE, pp. 1670-1673, 2021.

[24] A. Aminifar, F. Rabbi, V. K. I. Pun e Y. Lamo, «Monitoring Motor Activity Data for Detecting Patients' Depression Using Data Augmentation and Privacy-Preserving Distributed Learning,» 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society, pp. 2163-2169, 2021 .

[25] Z. Che, Y. Cheng, S. Zhai, Z. Sun e Y. Liu, «Boosting deep learning risk prediction with generative adversarial networks for electronic health records,» IEEE International Conference on Data Mining (ICDM), n. IEEE, pp. 787-792, 2017 .

[26] M. Ghalwash, Z. Yao, P. Chakrabotry, J. Codella e D. Sow, «ODVICE: An Ontology-Driven Visual Analytic Tool for Interactive Cohort Extraction,» arXiv preprint arXiv:2005.06434, 2020.

[27] C. Lu, B. Huang, K. Wang, J. M. Hernández-Lobato, K. Zhang e B. Schölkopf, «Sample-efficient reinforcement learning via counterfactual-based data augmentation» arXiv preprint arXiv:2012.09092, 2020.

[28] A. Massaro, V. Maritati, D. Giannone, D. Convertini e A. Galiano, «LSTM DSS automatism and dataset optimization for diabetes prediction,» Applied Sciences, vol. 17, n. 3532, p. 9, 2019.

[29] H. P. Das, R. Tran, J. Singh, X. Yue, G. Tison, A. Sangiovanni-Vincentelli e C. J. Spanos, «Conditional Synthetic Data Generation for Robust Machine Learning Applications with Limited Pandemic Data,» arXiv preprint arXiv:2109.06486, 2021.

[30] N. Sangari e Y. Qu, «A ComparMachine Learning Algorithms for Predicting Breast Cancer Prognosis in Improving Clinical Trials,» 2020 International Conference on Computational Science and Computational Intelligence (CSCI), pp. 813-818, 2020.

[31] A. Massaro, «Electronics in Advanced Research Industries: Industry 4.0 to Industry 5.0 Advances,» John Wiley & Sons, 2021.

[32] A. Massaro, «Human–Machine Interfaces,» Electronics in Advanced Research Industries: Industry 4.0 to Industry 5.0 Advances, vol. 10.1002/9781119716907.ch3, pp. 103-153, 2022.

[33] A. Massaro, G. Mastandrea, L. D'Oriano, G. R. Rana, N. Savino e A. Galiano, «Systems for an intelligent application of automated processes in industry: A case study from "PMI IoT Industry 4.0" project,» IEEE International Workshop on Metrology for Industry 4.0 & IoT, vol. IEEE, pp. 21-26, 2020.

[34] A. Massaro, G. Cosoli, V. Giardinelli, A. Leogrande, N. Magaletti, (2022). An Approach Generating Augmented Data For Machine Learning Data Processing, Zenodo. https://doi.org/10.5281/zenodo.6362964

Appendix A: hyperparameters adapted for the data processing and parameters calculated for the comparison

80-20 From the Top						
	Originals	**Plus or minus 3**	**Plus or minus 4**	**Plus or minus 5**	**Plus or minus 6**	**Plus or minus 7**
Mean Absolute Error	0,035535624	0,0170301	0,0176086	0,03366161	0,01	0,021
Mean Squared Error	0,001928981	0,0000000	0,0000000	0,00122413	0,000	0,001
Root Mean Squared Error	0,043920163	0,0191284	0,019957	0,03498753	0,02	0,024
	Absolute Variation					
Mean Absolute Error		-0,018505524	-0,017927024	-0,001874014	-0,025535624	-0,014535624
Mean Squared Error		-0,001928981	-0,001928981	-0,000704851	-0,001928981	-0,000928981
Root Mean Squared Error		-0,024791763	-0,023963163	-0,008932633	-0,023920163	-0,019920163
Average of Errors		-0,015075422	-0,014606389	-0,003837166	-0,017128256	-0,011794922
	Percentage Variation					
Mean Absolute Error		-52,07597807	-50,44803421	-5,273619302	-71,85922459	-40,90437164
Mean Squared Error		-100,0	-100,0	-36,5	-100,0	-48,2
Root Mean Squared Error		-56,44733822	-54,5607332	-20,33834191	-54,46282828	-45,35539393
Average of Errors		-69,5077721	-68,3362558	-20,71734041	-75,44068429	-44,80630512

Figure 9. Results of the prediction with the augmented data in comparison with the prediction with the original data in the case of partitioning 80-20 Take From the Top with indication of the statistical errors.

	80-20 Linear Sampling					
	Originals	**Plus or minus 3**	**Plus or minus 4**	**Plus or minus 5**	**Plus or minus 6**	**Plus or minus 7**
Mean Absolute Error	0,066452	0,040179228	0,041068725	0,044065677	0,042801352	0,042285963
Mean Squared Error	0,00814	0,004361432	0,004447278	0,004781614	0,004770372	0,004513238
Root Mean Squared Error	0,090222	0,066041141	0,066687915	0,069149215	0,069067877	0,067180636
	Absolute Variation					
Mean Absolute Error		-0,026272772	-0,025383275	-0,022386323	-0,023650648	-0,024166037
Mean Squared Error		-0,003778568	-0,003692722	-0,003358386	-0,003369628	-0,003626762
Root Mean Squared Error		-0,024180859	-0,023534085	-0,021072785	-0,021154123	-0,023041364
Average of Errors		-0,0180774	-0,017536694	-0,015605831	-0,016058133	-0,016944721
	Percentage Variation					
Mean Absolute Error		-39,53646542	-38,19790977	-33,68795973	-35,59057365	-36,36615452
Mean Squared Error		-46,4197543	-45,36513514	-41,25781327	-41,39592138	-44,55481572
Root Mean Squared Error		-26,80151072	-26,08464122	-23,35659263	-23,4467458	-25,53852054
Average of Errors		-37,58591015	-36,54922871	-32,76745521	-33,47774694	-35,48649693

Figure 10. Results of the prediction with the augmented data in comparison with the prediction with the original data in the case of partitioning 80-20 Take From the Top with indication of the statistical errors.

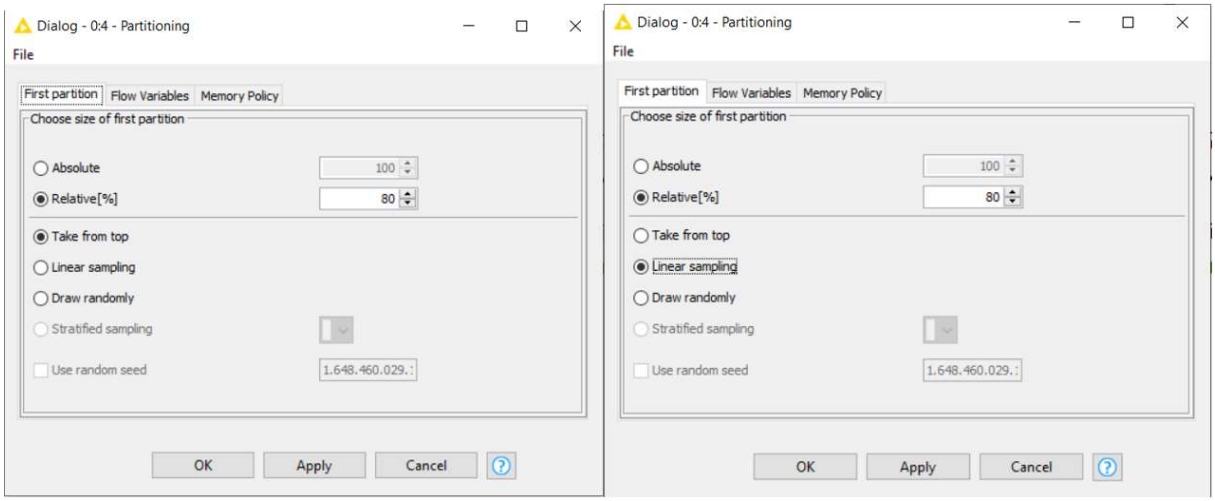

Figure 11. Hyperparameters of the partitioning node in KNIME to set the learning rate optimization methodologies for training the ANN-MLP algorithm.

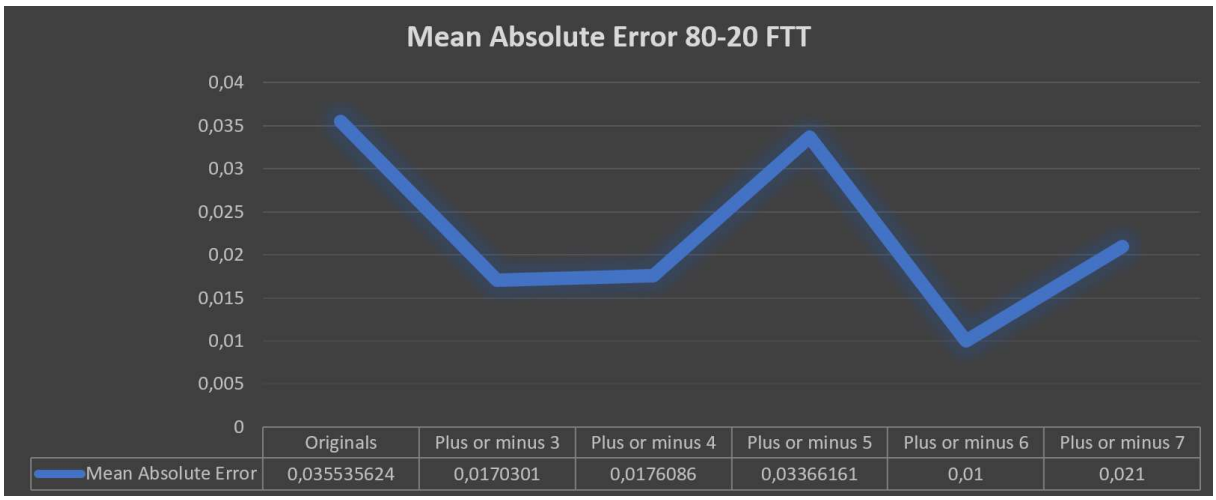

Figure 12. Trend of the Mean Absolute Error-MAE in the various configurations tested with the increased data. As evident from the graph it appears that the data increased with randomization around the mean for a value around + or - 6% are preferable in terms of minimization of the MAE in the case of Take from the Top.

Figure 13. Trend of the Mean Squared Error-MSE in the various configurations tested with the augmented data. As evident from the graph it appears that the data augmented with randomization around the mean for a value around + or - 6%, + o – 3%, and + o – 4% are preferable in terms of minimization of the MSE in the case of Take From The Top.

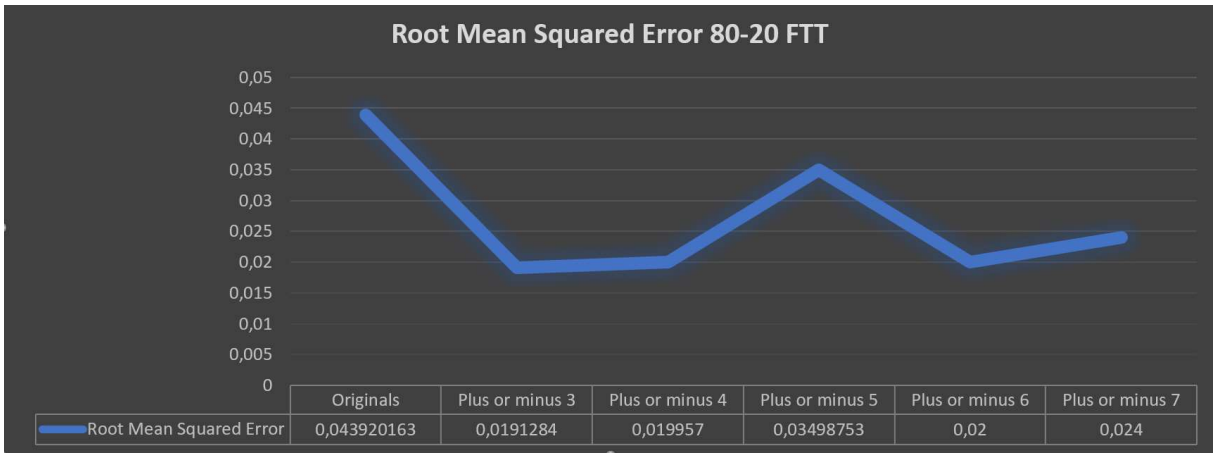

Figure 14. Trend of the Root Mean Squared Error-RMSE in the various configurations tested with the augmented data. As evident from the graph it appears that the data augmented with randomization around the mean for a value around + o – 3% is preferable in terms of minimization of the RMSE in the case of Take From The Top.

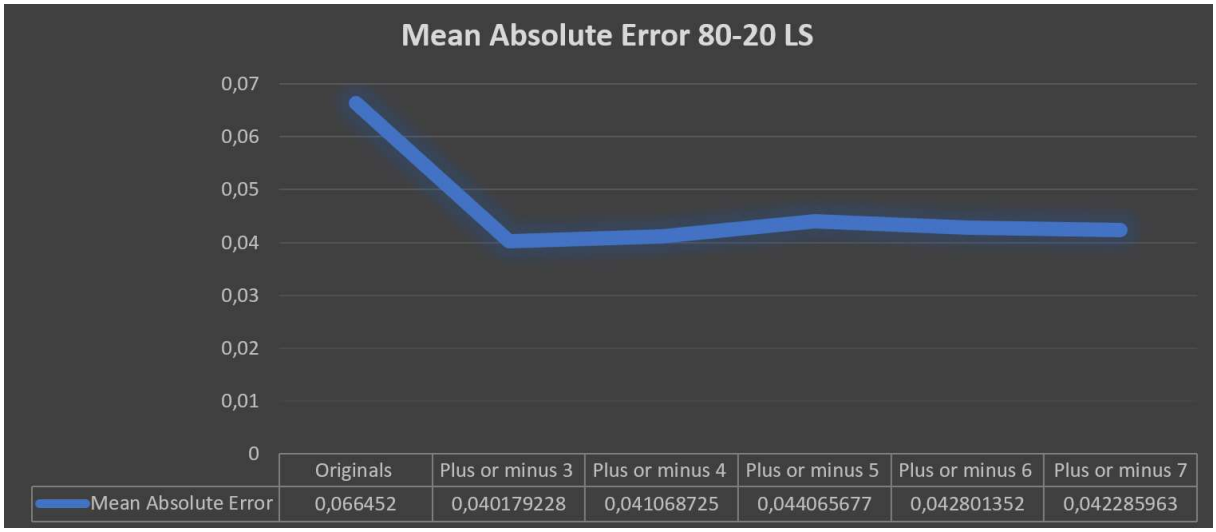

Figure 15. Trend of the Mean Absolute Error-MAE in the various configurations tested with the augmented data. As evident from the graph it appears that the data augmented with randomization around the mean for a value around + o – 3% is preferable in terms of minimization of the MAE in the case of Linear Sampling.

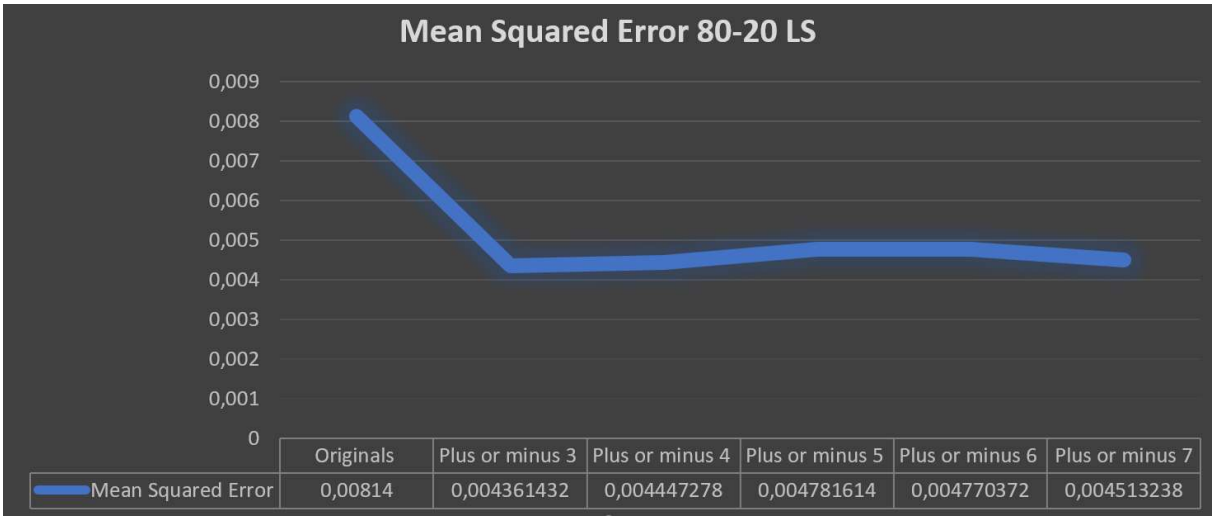

Figure 16. Trend of the Mean Squared Error-MSE in the various configurations tested with the augmented data. As evident from the graph it appears that the data augmented with randomization around the mean for a value around + o – 3% is preferable in terms of minimization of the MSE in the case of Linear Sampling.

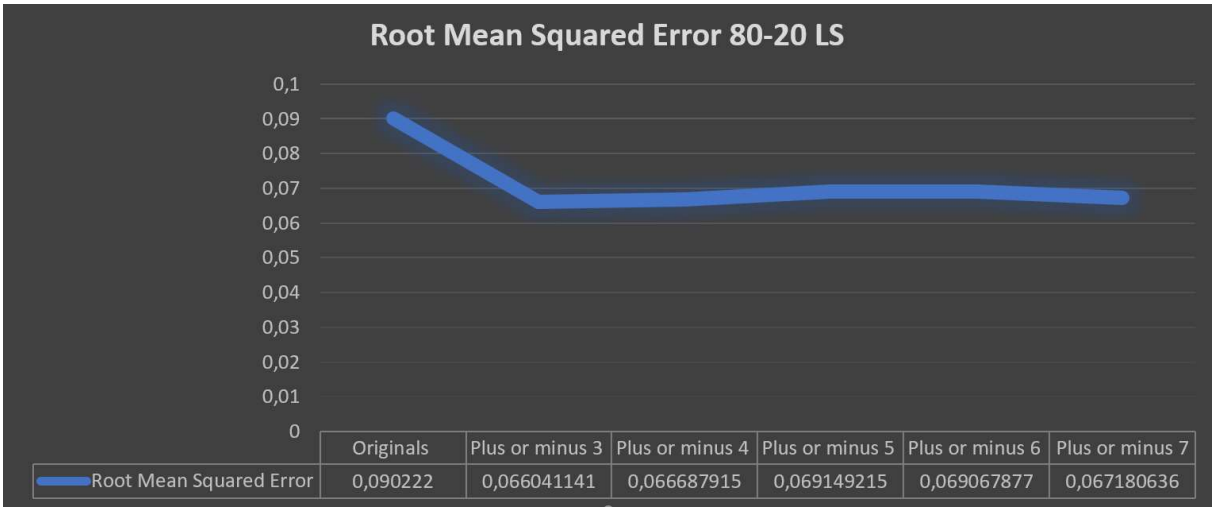

Figure 17. Trend of the Root Mean Squared Error-RMSE in the various configurations tested with the augmented data. As evident from the graph it appears that the data augmented with randomization around the mean for a value around + o – 3% is preferable in terms of minimization of RMSE in the case of linear sampling.