Techniques for Data Augmentation and Their Impact on Long-Range Dependence and Applications

MARYAM GHANBARI[1,2], WITOLD KINSNER[1], NARIMAN SEPEHRI[2]
[1]Electrical and Computer Engineering Department,
University of Manitoba,
Winnipeg, MB,
CANADA

[2]Mechanical Engineering Department,
University of Manitoba,
Winnipeg, MB,
CANADA

Abstract: - Data augmentation is a common approach to enhance datasets for training machine learning models. This study employs five distinct techniques to generate augmented datasets. Furthermore, eight measures are applied to assess datasets both before and after augmentation techniques. A critical requirement is that any augmentation should preserve the fundamental properties of the original dataset. The study reveals that certain augmentation methods can disrupt the long-range dependence on *Internet traffic data* (ITD) *with distributed denial of service* (DDoS) *attacks* (DDoS ITD). These DDoS ITDs originate from stochastic and bursty environments, affecting the *probability mass function* (PMF) and data labeling.

Key-Words: - Long-range dependence, data augmentation, Internet traffic data, speech sound data, probability mass function, dataset evaluation measures.

Received: April 9, 2024. Revised: October 11, 2024. Accepted: November 15, 2024. Published: December 23, 2024.

1 Introduction

This paper presents a study of data augmentation when it is applied to processes with long-range dependence to show that some augmentation techniques may alter or even destroy the long-range dependence. This paper is an extension of our previous work presented at the IEEE International Conference on Cognitive Informatics and Cognitive Computing [1] to develop the foundation for our research in this area. In this extended paper, we delve deeper into the applications of long-range dependence and include a more comprehensive set of case studies.

Data augmentation is a commonly employed method for enlarging and diversifying datasets during the training of neural networks. The advantage of data augmentation has been demonstrated across various domains, including image classification and data modeling, facilitating the expansion of training samples, [2]. Bjerrum *et al.* applied *extended multiplicative scattering* (EMSC) to correct the datasets, as well as a *spectral data augmentation method* to augment the datasets using random variations in slope and offset, [3]. They used convolutional neural networks to extract features from the data, and they demonstrated that the combination of data augmentation and EMSC was the best preprocessing method to enhance test results.

A data augmentation technique is correct and useful when the constraints of the dataset are preserved. For example, by augmenting an image dataset or an audio dataset, the properties of the original and augmented dataset remain the same, [2]. The principal requirement of any data augmentation technique is that it should not alter the constraints of a dataset. For example, when an augmentation technique modifies the data *probability mass function* (PMF) of an Internet dataset, the properties of the dataset are modified. Therefore, since the long-range dependence of the Internet dataset is altered, the augmentation technique is unsuitable.

This paper examines data augmentation for two distinct categories of datasets: (i) Internet traffic data with *distributed denial of service* (DDoS) attacks and (ii) the Manitoba Speech Dataset with standard voice sound collection, [1]. The study

demonstrates that the speech dataset constraints can be preserved using the selected data augmentation methods, but these methods cannot preserve the constraints of the Internet traffic data containing the DDoS attacks (DDoS ITD), given its stochastic nature. Techniques such as adding noise, mirroring, squeezing, and expanding the DDoS ITD alter the data's shape and PMF, indicating that conventional augmentation methods may not be suitable for DDoS ITD. Mono-scale, multi-scale, and poly-scale measures are employed to assess the sensitivity of these techniques to various factors by analyzing the time series datasets before and after augmentation.

The structure of this paper is outlined as follows. Section 2 provides a summary of the datasets. Section 3 explains the methods of data augmentation. Section 4 introduces the measures utilized for dataset analysis. Sections 5 and 6 present the simulation outcomes and provide an explanation and discussion of the results, respectively. Section 7 discusses the potential applications. Section 8 introduces an extension of the research. Section 9 offers concluding remarks.

2 Description of the Datasets

This study uses two types of time series data sets. Each complete dataset is referred to as an epoch (T_E), with individual stationary segments within each epoch designated as a frame. The word "window" is not used in this paper as it is reserved to signify frame data modifications at the edges of the frame through a window function (also known as a tapering or apodization function) such as the Hamming and Hann windows, often employed in signal processing.

In this paper, we consider only discrete signals that have been sampled from continuous (analog) signals properly to represent the original signal. The sampling frequency, f_s, is adequate if it is strictly above the Nyquist frequency, $f_s > f_N = 2f_c$. For narrowband signals, the critical cutoff frequency, f_c, is defined at the 3-dB drop in the log-log plot of the frequency response of the signal. For broadband signals originating from many self-affine processes, the frequency is defined at a higher value, $f_c = f_h$, where the log-log response reaches the noise level of the signal.

2.1 Internet Traffic Data

The dataset employed for training and testing purposes in this study was sourced from the trusted Center for *Applied Internet Data Analysis* (CAIDA). CAIDA collects a diverse range of real-time network traffic from various parts of the world in collaboration with research organizations, governments, and commercial entities without revealing their identities, [4].

The CAIDA's 2007 DDoS attack traffic was used as network traffic data and contains TCP, UDP Flood, SYN Flood, and ICMP (Ping) Flood packets. A packet in network traffic data has the following features: source IP address, destination IP address, packet arrival time, packet length, and protocol. The critical attribute called the packet arrival time series signal is computed by calculating the difference between a packet's arrival time at *t* and its arrival time at *t*-1. The distribution of packets with a duration of 0.1 *ms* within each frame is illustrated in Figure 1.

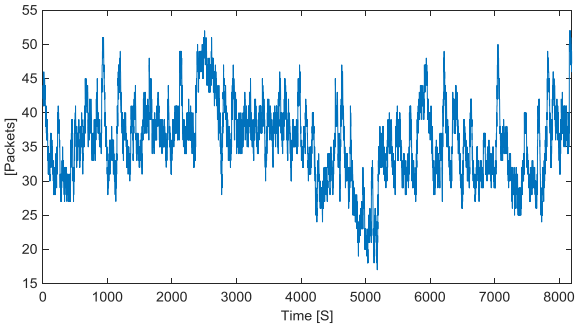

Fig. 1: The number of packets within a stationary frame size

2.2 Speech Sound Data

For comparison with the CAIDA dataset, the Manitoba Speech Dataset was utilized, obtained from the University of Manitoba, [1]. This dataset contains recordings of 44 words spoken by 12 female and 12 male volunteers. This study uses the word "test" epoch as a speech sound dataset. Figure 2 illustrates the plot of the "test" dataset, sampled at a rate of 44.1 *kilo samples per second* (kSps).

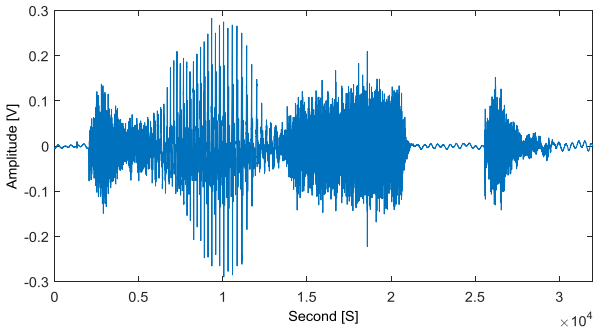

Fig. 2: Plot of the dataset for the word "test"

3 Data Augmentation Techniques

This study employs the following five distinct techniques to generate six augmented sets.

3.1 Mirroring Technique

The mirroring technique, also known as flipping, involves reversing the order of the time series data within a stationary frame. In this form of horizontal flipping, the first sample in the series becomes the last, while the last sample becomes the first. This transformation effectively reverses the temporal order, introducing a new perspective of the data without altering the underlying magnitude or amplitude patterns.

3.2 Time Stretching Technique

Time stretching is a resampling method that aims to alter the temporal resolution of the time series data. In this technique, the original time series is upsampled by inserting additional data points between existing samples. The new points are generated by calculating the average of neighboring values and placing them at even positions within the data. Meanwhile, the existing points are shifted to the right, increasing the overall length of the time series. As illustrated in Figure 3, time stretching preserves the general shape and trends of the original signal.

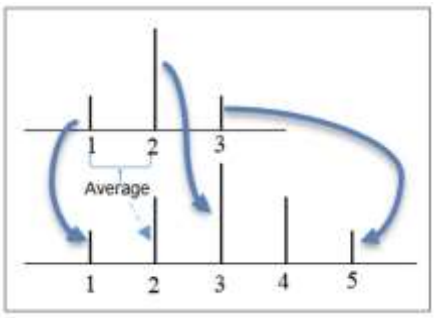

Fig. 3: The upsampling technique applied to a time series dataset

3.3 Squeezing Technique

The squeezing technique reduces the data resolution by removing samples to downsample the time series. It is an effective approach for minimizing the size of the dataset while retaining key patterns and trends. This technique can be accomplished in the following two ways:

3.3.1 Downsampling

This technique selects data points located at odd positions within the time series, as illustrated in Figure 4.

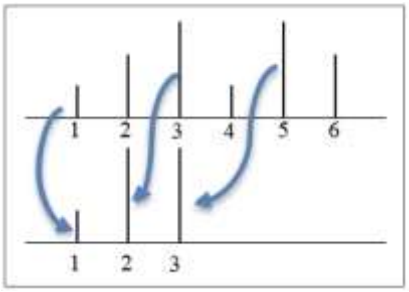

Fig. 4: Downsampling technique applied to a time series dataset, [1]

3.3.2 Wavelet Approximation Coefficients

Using the *Daubechies wavelet transform 2* (db2) basis function, this technique downsamples the data by extracting approximate coefficients. These coefficients are then downsampled within a stationary frame, as illustrated in Figure 5.

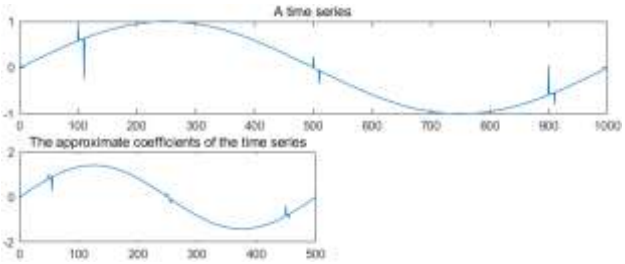

Fig. 5: Downsampling technique applied to a time series dataset using the approximate coefficients of the dataset

3.4 Random Cut-and-Paste Technique

The random cut-and-paste technique generates augmented data by randomly selecting a segment of the original time series and appending it to the end. This operation disrupts the natural order of the time series, creating a synthetic sequence that may combine different temporal patterns. By altering the original data flow, this technique introduces new transitions and relationships between segments, challenging the model to learn more complex temporal dependencies. The cut-and-paste approach is useful when the dataset is limited, as it can create a wide range of augmented samples from a single time series.

3.5 Adding White Noise Technique

The addition of white noise involves injecting a random noise component into the time series, effectively creating a perturbed version of the original data. The white noise is characterized by a normal distribution with a mean of zero and a small variance, typically set to 0.0001, to ensure minimal distortion. This technique simulates random

fluctuations and measurement errors that may occur in real-world data.

4 Selecting Measures for the Analysis of Datasets

In order to assess the sensitivity of the five techniques to various factors, seven mono-scale, multi-scale, and poly-scale measures are employed to analyze the time series of the two datasets as described below.

4.1 The Hurst Exponent Measure

The *Hurst exponent* (*H*) serves as a metric (measure) for quantifying the degree of long-range dependence present in a time series [5], thereby detecting its existence. Additionally, the Hurst exponent measures the smoothness of self-affine processes. It ranges between 0 and 1, where different values convey distinct characteristics of the time series, [5]. As the *H* value approaches 1, the level of persistence or *long-range dependence* (LRD) increases. This implies that the signal's behavior at a given time can be influenced by its past values, a concept explained later in this paper. A Hurst exponent close to 0.5 suggests a completely random process or Brownian motion where there is no LRD. In this case, the time series exhibits no correlation, and the past values do not influence future data points. Such behavior is typical of white noise and purely stochastic processes, [5]. On the other hand, an *H* value less than 0.5 indicates an anti-persistence or strong negative correlation. In this scenario, if the time series has an increasing trend, it is likely to reverse and start decreasing, and vice versa.

An *N*-dimensional object (a 2D signal in our case) is said to be self-affine if its smaller fragments are scaled-down versions of the entire object and if the scaling factors are different in the *N* dimensions. If the scaling factors are the same, the object is called self-similar, [6]. Self-affinity refers to the persistence of fractal patterns across various scales of observation, while self-similarity denotes uniform scaling behavior across all dimensions.

Self-affinity and long-range dependence in a signal means that the signal not only exhibits fractal patterns at different scales but also these patterns are strongly correlated (or dependent) over an extended period of time. These relationships can be encountered in various real-world phenomena, such as financial time series and Internet traffic, where bursty operations and packet flows display similar behavior. The ability to detect and measure these patterns using the Hurst exponent is crucial for analyzing the predictability and underlying dynamics of time series data.

4.2 The Variance Fractal Dimension Trajectory Measure

The variance (the second moment) of a self-affine signal can be used to measure the power-law behavior of the signal. To measure and analyze the complexity of such broadband self-affine time series signals (often characterized by the signal stochasticity, non-stationarity, non-differentiability, dynamic behavior, and long-range dependence), the *variance fractal dimension trajectory* (VFDT) measure has been proposed, [6]. The VFDT has been refined to consider all data points, including boundary points of time series signals, not solely marginal points, [6].

When using the VFD with a time series, the variance of the amplitude increments over each time increment follows a power law relation with that time increment, as given by [6].

$$\mathrm{var}[A(t_2) - A(t_1)] \sim | t_2 - t_1 |^{2H} \tag{1}$$

where *var* denotes the variance function (the second moment), *A* is the time series signal, *H* is the Hurst exponent, and *it* represents the discrete time in a discrete signal. As explained, the Hurst exponent indicates the self-affine characteristics of the signal, if any. Taking the logarithm of both sides of Eq. (1), the Hurst exponent is calculated from a log-log plot and is given by [6].

$$H = \lim_{\Delta n \to 0} \frac{1}{2} \left(\frac{\log_2[\mathrm{var}(\Delta A)_{\Delta n}]}{\log_2 \Delta n} \right) \tag{2}$$

where Δn denotes the scale in a discrete-time series signal at which the variance is evaluated. Equation (2) is used to analyze a time series signal in the time domain by calculating the expansion of the time series signal amplitude at different scales through its variance [6].

The output of the VFD, denoted by D_σ is used as a measure of signal complexity and is obtained from [6].

$$D_\sigma = E + 1 - H \tag{3}$$

where *E* is the Euclidean dimension and is *E = 1* for time series signals. Detailed descriptions of the real-time VFDT algorithm and the second version of the VFDT algorithm can be found in [6].

To validate the VFDT algorithm, the theoretical VFDT value for white noise is D_σ =2, while D_σ =1 for a straight line, [6]. Thus, the VFDT values of a

self-affine time series signal are bounded between the two values. An example of the engineering application of the *H* and the VFD is to facilitate anomaly detection in Internet time series traffic.

4.3 Spectral Fractal Dimension Measure

The *spectral fractal dimension* (SFD) transforms a self-affine time series self-affine signal into its power spectrum density [6]. Operating in the frequency domain [1], the SFD analyzes the properties of frequencies underlying the time series signals, [6]. This spectrum discovers inherent characteristics of the time series frequencies [6]. For the fractal time series signal, the power spectrum density follows the spectral power law, as shown in Eq. (4), [6].

$$P(f) \sim \frac{1}{f^{\beta}} \tag{4}$$

where β is the power spectrum exponent. Taking the logarithm of both sides of Eq. (4) yields a linear relationship between the power spectrum and its frequency, represented by a regression line. The slope of this line, denoted by β, serves as the power spectrum exponent of the DDoS ITD, as illustrated in Figure 6.

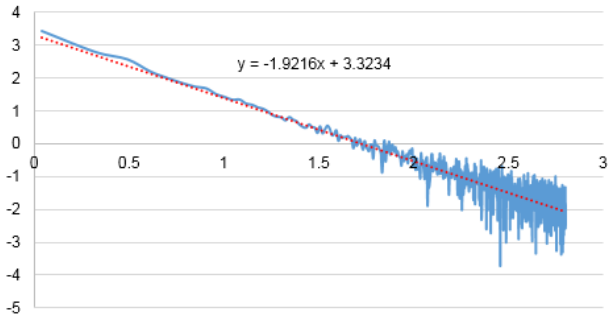

Fig. 6: The power density spectrum of the DDoS ITD and its slope

In Figure 6, the power density spectrum of the DDoS ITD is plotted on a log-log scale. The figure demonstrates the linear trend in the log-log plot, with the slope β representing the power spectrum exponent of the DDoS traffic. This exponent signifies the self-affine relationship of the time series. A slope of zero (β=0) indicates white noise, suggesting no correlation within the time series and implying purely random behavior. Consequently, an increasing slope indicates an increasing correlation. The analysis of the power density spectrum in Figure 6 indicates strong long-range dependence and persistent behavior. This observation is consistent with the nature of DDoS attacks, where packet bursts and correlated traffic patterns are common due to the nature of the DDoS attack.

Complex self-affine time series, such as the Internet time series data, may exhibit multiple β values due to their multifractal nature. The SFD output, serving as a measure of signal complexity denoted by D_{β}, is obtained from [6].

$$D_{\beta} = E + \frac{(3-\beta)}{2} \tag{5}$$

where E is the Euclidean embedding dimension, as defined in Eq. (3). This is due to the following relationship between the Hurst exponent and the power spectrum exponent [6].

$$\beta = 2H + 1 \tag{6}$$

4.4 Autocorrelation Function Measure

Correlation assesses both the strength and direction of the linear relationship between two variables, [7]. Autocorrelation, on the other hand, approximates the similarity between data points separated by successive time intervals within a time series signal, [8]. Equation (7a) presents the autocorrelation function for an energy signal, while Eq. (7b) shows the autocorrelation function for a power signal.

$$r_{xx}(l) = \sum_{n=-\infty}^{+\infty} x(n)x(n+l) \tag{7a}$$

$$r_{xx}(l) = \lim_{M\to\infty} \frac{1}{2M+1} \sum_{n=-M}^{M} x(n)x(n+l) \tag{7b}$$

Autocorrelation reveals properties of a stationary random process, [9]. The Fourier transform is computed from a stationary time series signal, allowing for the direct calculation of the power density spectrum from the squared magnitude of the Fourier transform. In the case of non-stationary data, the Fourier transform can still be calculated from a sequence of stationary frames. The Fourier transform of the autocorrelation sequence is generally valid for non-stationary data, enabling the calculation of the power of the time series. Consequently, the Fourier transform of the autocorrelation sequence is interpreted as the frequency distribution of the signal's power, representing the power density spectrum, [9]. For example, analyzing autocorrelation in voice datasets aids in pitch detection, while analyzing Internet traffic flow datasets facilitates anomaly detection.

4.5 Long-Range Dependence Measure

In statistical analysis, dependence signifies any relation and association between two variables. LRD describes the extent to which a time series signal is influenced by its past values over an

extended period. Unlike short-term dependence, where correlations diminish quickly, LRD implies that significant correlations persist even as the time lag increases, indicating memory and persistence within the time series, [10], [11]. The magnitude of long-range dependence is typically measured using the autocorrelation function, which measures the correlation of a time series with its lagged versions.

In instances where a time series signal exhibits long-range dependence, its autocorrelation decays in a hyperbolic manner [11], whereas signals with short-term dependence experience exponential decay [12]. Consequently, the autocorrelation distribution of a time series signal with long-range dependence tends to be more dispersed compared to that of a time series with short-term dependence [11], [13]. LRD is a key indicator of the effectiveness of data augmentation. Because the autocorrelation function decays hyperbolically, the power density spectrum for a signal with long-term dependence shows an increase without the frequency ever reaching zero, [11].

Long-range dependence indicates that past events can influence future events over an extended period, [14]. For example, Internet traffic patterns, particularly in scenarios like DDoS ITD, tend to be bursty and exhibit LRD [14]. Another example is Call Holding Time (CHT), which occurs in a stochastic (random) environment and follows a heavy-tailed distribution while also showing LRD. In telecommunications, CHT, or the duration of a phone call, varies randomly due to factors such as user behavior, network conditions, and the type of call. Processes characterized by long-range dependence, such as CHT and Internet traffic, often exhibit self-affinity. As a result of this LRD, periods of high activity are likely to be followed by similar high activity periods, while low activity periods tend to be followed by low activity periods. Both CHT and Internet traffic follow a heavy-tailed distribution, meaning that extreme values (such as unusually very long or very short call durations or traffic levels) are common and support the presence of bursty patterns.

The power spectrum exponent indicates the presence of long-range dependence within a time series signal [11] and can be determined using Eq. (6). The time series represents white noise when the slope β is zero, indicating no correlation and, therefore, no long-range dependence. As the slope increases, so does the correlation, subsequently enhancing the long-range dependence of the time series signal.

4.6 Zero Crossing (ZC) Measure

The zero crossing (ZC) measure quantifies how many times a signal's magnitude crosses a specified threshold value, such as zero, within a given interval [1]. When the threshold value is set to zero, the ZC indicates the rate of transitions between positive and negative mathematical signs of the signal within that interval. Zero crossing is effective in identifying edges and sudden changes within a time series signal and is related to the lowest frequency component of the signal [15], which can be utilized for feature extraction. The zero crossing is calculated through the following equation [1], [16].

$$ZC = \sum_{n=-\infty}^{\infty} [|\,\mathrm{sgn}[x(m)] - \mathrm{sgn}[x(m-1)]\,|]w[n-m] \quad (8)$$

where $\mathrm{sgn}[]$ denotes the mathematical sign function as given by [1], [16].

$$\mathrm{sgn}[x(n)] = \begin{cases} +1, & x(n) \geq threshold \\ -1, & x(n) < threshold \end{cases} \quad (9)$$

and $w[]$ denotes a frame containing a stationary segment of a time series signal, defined by [16].

$$w[n] = \begin{cases} \dfrac{1}{2N}, & 0 \leq n \leq N-1 \\ 0, & otherwise \end{cases} \quad (10)$$

This measure is computed in the time domain and can be performed in real-time. The ZC rate is useful for distinguishing speech from noise and for determining the start and the end of speech segments, [16].

4.7 Turns Count (TC) Measure

The turns count (TC) is used to extract features from a time series signal based on changes in slope direction rather than zero crossings, [15]. A turn is counted each time the slope of the signal changes its sign, [1], [15]. This technique evaluates signals by identifying the number of spikes present within the signal [15] and is related to the highest frequency component of the signal in the frame. A turn is computed from [1].

$$tr_i = x(n) > x(n+1) \,\&\, x(n+1) < x(n+2) \quad (11a)$$

$$tr_i = x(n) < x(n+1) \,\&\, x(n+1) > x(n+2) \quad (11b)$$

where tr_i represents the turn occurring within a time interval such as a stationary-frame interval, and *x(n)* denotes an input time series signal. The total turn count in the frame is given by [1].

$$TC = \sum_{i=1}^{\infty} tr_i \quad (12)$$

5 Simulation and Results

Table 1 and Table 2in Appendix present the statistical properties and outcomes related to eight mono-scale, multi-scale, and poly-scale assessments of both the DDoS ITD and the "test" dataset before and after the implemented data augmentation. Subsequent discussions will explain the findings of the simulated measure proposed.

5.1 Analysis of Stationarity

Weak stationarity in a dataset is characterized by the first two moments falling within specified ranges, typically within a 95% confidence interval, [6]. The DDoS ITD demonstrated weak stationarity, as indicated by the trajectory of the mean falling within the range of 0.000782 to 0.003041 and the variance trajectory within the range of 1.64E–06 to 8.21E–06, given a minimum frame size of 8192. The skewness trajectory ranged from 1.076 to 2.94, while the kurtosis trajectory remained at 1.8. The trajectories of the mean and variance are presented in Figure 7, while the trajectories of skewness and kurtosis are illustrated in Figure 8.

Recall that the mean represents the typical value of a set of data, while variance quantifies how spread out the data is from this mean, [7]. Skewness signals if the distribution leans towards the right or left, [17]. Kurtosis measures the thickness of the distribution's tail, with positive values indicating distributions with heavier tails.

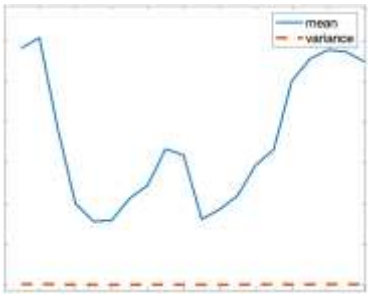

Fig. 7: The trajectories of the mean and variance within a stationary frame of samples related to the DDoS ITD

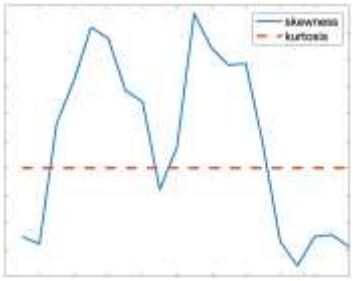

Fig. 8: The trajectories of skewness and kurtosis within a stationary frame of samples related to the DDoS ITD

The "test" speech dataset exhibited weak stationarity due to the minimum window size of 512, resulting in the mean trajectory ranging from–0.0087 to 0.0049 and the variance trajectory ranging from 2.48E–06 to 0.0091. Similarly, the skewness trajectory ranged from –2.09 to 1.55, and the kurtosis trajectory ranged from 1.56 to 31.03. The trajectories of the mean and variance are presented in Figure 9, while the trajectories of skewness and kurtosis are illustrated in Figure 10.

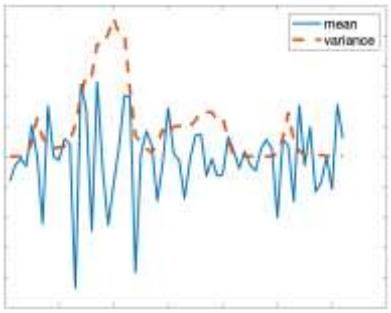

Fig. 9: The trajectories of mean and variance within a stationary frame of samples related to the "test" dataset

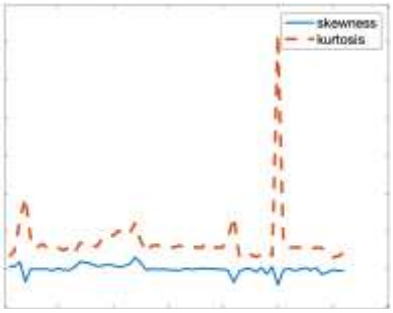

Fig. 10: The trajectories of skewness and kurtosis within a stationary frame of samples related to the "test" dataset

5.2 Probability Mass Function Analysis

To fit a probability distribution to the time series of both datasets, the maximum likelihood estimation method and polynomial regression are employed, [18].

For the DDoS ITD, the probability mass function (PMF) reveals a Lévy distribution, where the location parameter (μ) is 0.0019223631, and the scale parameter (σ) is –0.0004226877 as given by:

$$f(x) = \sqrt{\frac{\sigma}{2\pi}} \cdot \frac{e^{-\frac{\sigma}{2(x-\mu)}}}{(x-\mu)^{3/2}} \quad x > \mu, -\infty < \mu, \sigma > 0 \qquad (13)$$

These parameters are determined through maximum likelihood estimation. The DDoS ITD's heavy tail becomes apparent when contrasting its Lévy distribution with a normal distribution. The

presence of self-affinity in the delay of packets contributes to the heavy tail distribution, [13]. The on-off queuing model, where "on" represents active data transmission and "off" represents no transmission, generates the traffic model. If the durations of "on-off" periods exhibit heavy-tailed characteristics, it leads to long-range dependence, [5]. Figure 11 illustrates the PMF of the DDoS ITD.

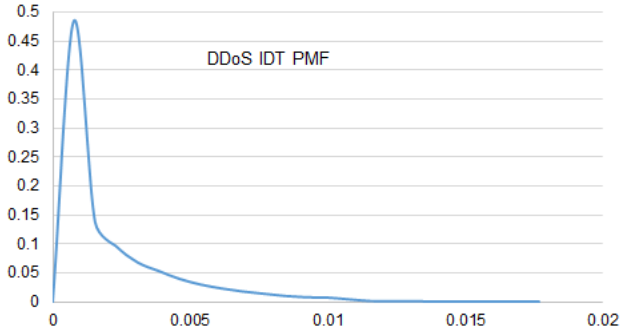

Fig. 11: The Probability Mass Function (PMF) of the DDoS ITD

The PMF of the "test" speech dataset conforms to a distorted (skewed and flattened) normal distribution, characterized by a mean (μ) of 0.090909091 and a standard deviation (σ) of 0.028493958. These values are obtained through maximum likelihood estimation. Figure 12 illustrates the PMF of the "test" dataset.

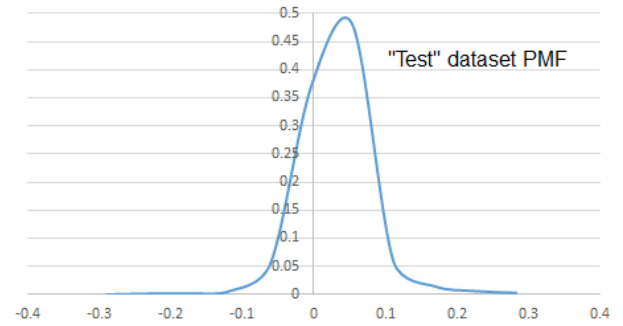

Fig. 12: The Probability Mass Function (PMF) of the "test" dataset

5.3 The Hurst Exponent Analysis

In this study, the DDoS ITD exhibited Hurst exponent (2) values ranging from 0.0063 to 0.0289, as illustrated in Figure 13. Similarly, the "test" dataset demonstrated Hurst exponent values ranging from 0.0137 to 0.8520, as illustrated in Figure 14.

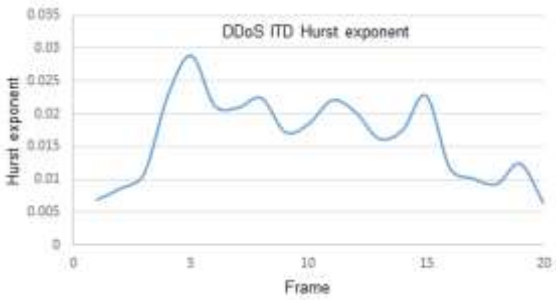

Fig. 13: The trajectory of the Hurst exponent for the DDoS ITD

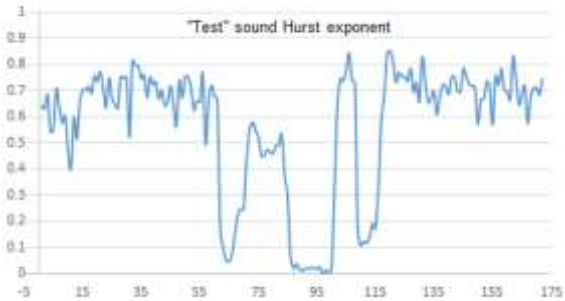

Fig. 14: The trajectory of the Hurst exponent for the "test" dataset

5.4 Variance Fractal Dimension Trajectory Analysis

The VFD algorithm output for the DDoS ITD ranged from 1.9711 to 1.9937 using the non-overlapping frame version, as shown in Figure 15. Similarly, for the "test" dataset, the VFD ranged from 1.1480 to 1.9863, as illustrated in Figure 16.

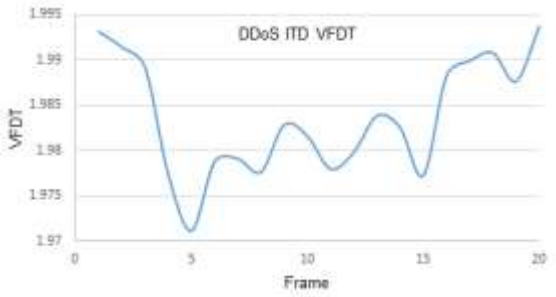

Fig. 15: The trajectory of the VFD for the DDoS ITD

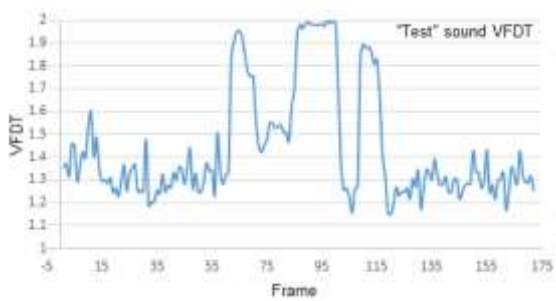

Fig. 16: The trajectory of the VFD for the "test" dataset

To validate the VFD algorithm results, a uniformly distributed white-noise time series with a mean of zero, variance of 0.08, and an epoch size of 220 was generated. This white noise within the same frame of 512 samples of stationary time series data is shown in Figure 17.

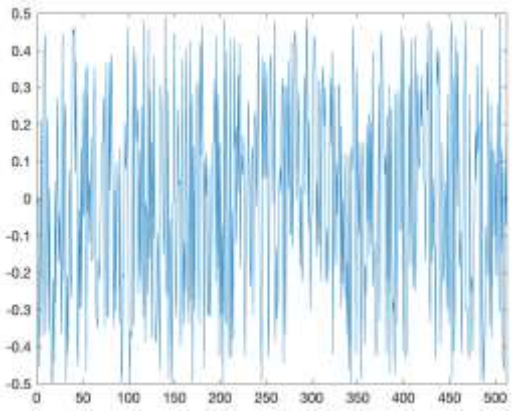

Fig. 17: Uniformly distributed white noise is illustrated within a 512-element frame

For the white noise epoch, the VFD output using the non-overlapping frame version ranged from 1.955 to 1.995, with an absolute error of –0.0241 and an absolute error of –4.15%, confirming the accuracy of the algorithm. Figure 18 illustrates the VFDT of this white noise signal with zero overlapping frames for the epoch.

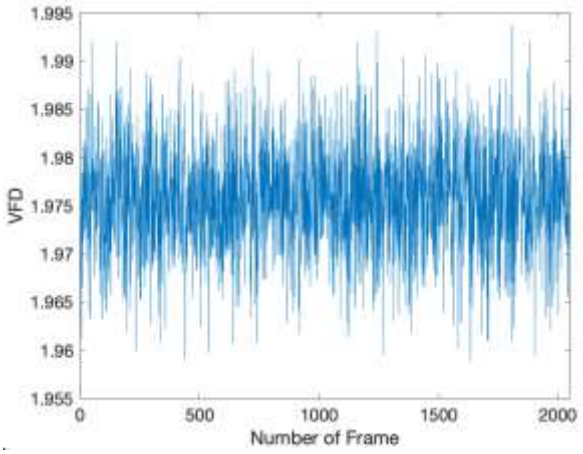

Fig. 18: The VFDT of the uniform distribution white noise signal

5.5 Zero Crossing and Turns Count Analysis

In this study, the DDoS ITD shows 1250 zero-crossing, and 5349 turns count within a frame size of 8192, while the "test" dataset exhibits 145 zero-crossing and 192 turns count within a frame size of 512.

5.6 Power Density Spectrum Analysis

The power density spectrum of both the DDoS ITD and the "test" datasets are illustrated as log-log plots in Figure 19 and Figure 20, respectively. The power density spectrum has a slope of 1.9216 for the DDoS ITD and 1.4727 for the "test" dataset when plotted against frequency. However, according to Eq. (6), the power spectrum exponent β is 0.9728 for the DDoS ITD and 1.6095 for the "test" dataset, respectively.

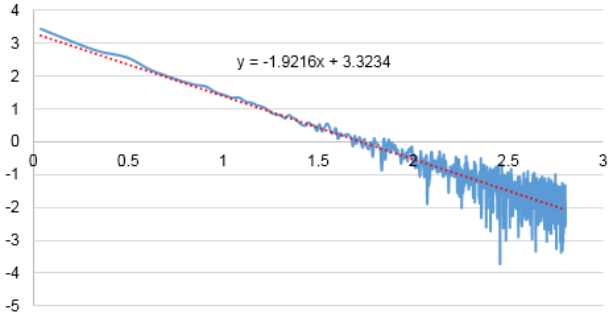

Fig. 19: A log-log plot showing the power density spectrum against frequency, along with the corresponding slope for the DDoS ITD

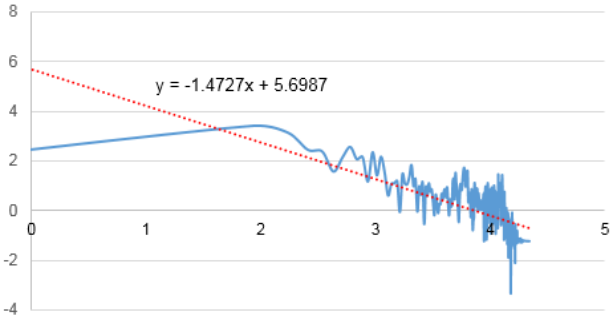

Fig. 20: A log-log plot showing the power density spectrum against frequency, along with the corresponding slope for the "test" dataset

5.7 Autocorrelation Function Analysis

The autocorrelation for the DDoS ITD and the "test" dataset is illustrated in Figure 21 and Figure 22, respectively.

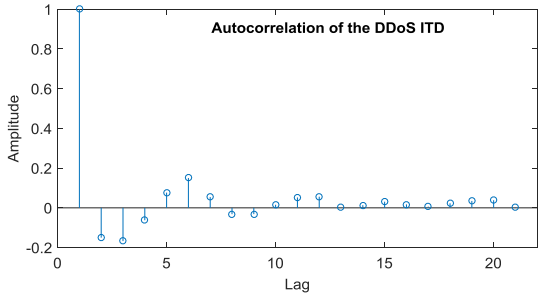

Fig. 21: The DDoS ITD's autocorrelation at different lags in the first frame

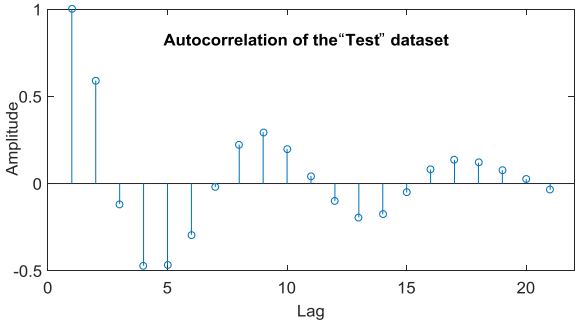

Fig. 22: The "test" dataset's autocorrelation at different lags in the first frame

6 Results and Discussion

This study reveals that data augmentation affects the constraints of the DDoS ITD, whereas the constraints of the "test" dataset remained unchanged.

Typically, a Lévy distribution exhibits infinite mean and variance with undefined skewness and kurtosis. Also, in a Lévy distribution, inflection points vary in magnitude, making it impossible to determine the sigma to find the variance. Consequently, these statistical tools are inadequate for feature extraction. However, the PMF serves as a convenient statistical tool for displaying data distribution to check whether the shape of the data has changed as a result of the augmentation.

In this study, the PMF of the DDoS ITD followed a Lévy distribution with finite values for its first four moments. Although these values were finite, they were appropriate since the DDoS ITD data had a limited stationary frame size, ensuring the data magnitude within this frame remained finite.

The augmentation techniques that violate the fundamental characteristics of the data also violate the constraints of the time series signals. However, the constraints are altered when the augmented DDoS ITD does not conform to the Lévy distribution. For example, employing the stretching technique for data augmentation results in a polynomial distribution in the PMF of the DDoS ITD. Similarly, utilizing the squeezing technique leads to a Pareto distribution in the PMF, while employing the random cut-and-paste technique results in a Lévy distribution. These distributions are illustrated in Figure 23, Figure 24, Figure 25, Figure 26, and Figure 27, respectively.

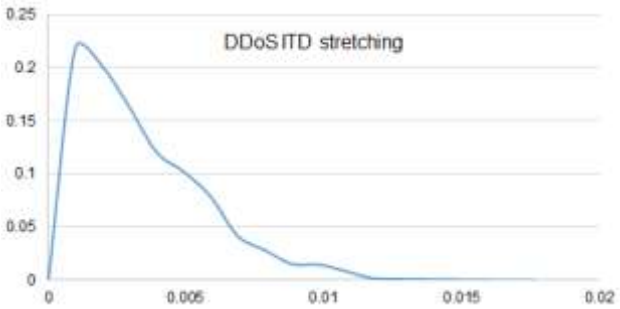

Fig. 23: After applying the stretching technique, the PMF of the augmented DDoS ITD results in a polynomial of degree 9 distribution. The coefficients for this polynomial are [4.7e-09, –4.5e-07, 1.8e-05, –0.0004, 0.0062, –0.0548, 0.3054, –1.0178, 1.7902, –1.0285]

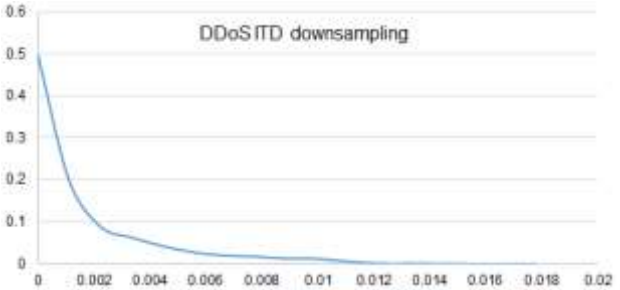

Fig. 24: After applying the squeezing technique (downsampling), the PMF of the augmented DDoS ITD results in a Pareto distribution with parameters: Shape equal to 1.71292, Scale equal to 0.00601753 and Threshold equal to 0.

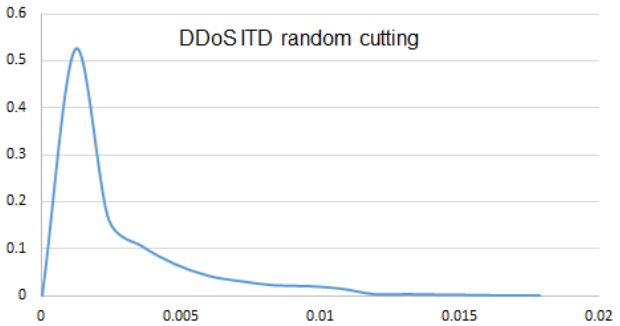

Fig. 25: After applying the random cut-and-paste technique, the PMF of the augmented DDoS ITD results in a Lévy distribution with parameters μ equal to 0.07692308 and σ equal to 0.006554275

Fig. 26: After applying the mirroring flipping horizontally technique, the PMF of the augmented DDoS ITD results in a Lévy distribution with parameters μ equal to 0.0588451 and σ equal to 0.00346275

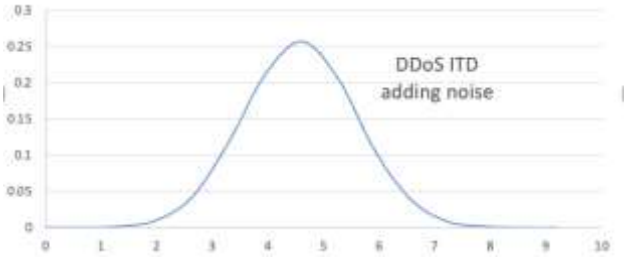

Fig. 27: After applying the adding noise technique, the PMF of the augmented DDoS ITD results in a normal distribution with parameters μ equal to 0.067 and σ equal to 0.008234

Flipping and random cut-and-paste of the DDoS ITD resulted in similar PMF patterns, whereas stretching, squeezing, and adding noise did not yield similar outcomes. With the horizontal flipping technique, the tagging (labeling) of normal or anomalous data output cannot be maintained since the tag can be altered. Data output was labeled according to the packet count within a 0.1 ms interval, with fewer than 40 packets considered normal and more than 40 packets considered anomalous. The horizontal flipping technique changed the data output tag and made this augmentation technique impractical for machine learning model training, [2]. Moreover, the random cut-and-paste technique is not suitable for adequately expanding small datasets due to their non-stationary nature; thus, it is only viable for large datasets where augmentation is unnecessary due to a large number of data points. Furthermore, the data output tag can be modified.

The PMF distributions of the "test" dataset augmented with stretching, squeezing, and random cut-and-paste techniques result in normal distributions, as illustrated in Figure 28, Figure 29, Figure 30, Figure 31 and Figure 32, respectively.

In voice datasets, flipping, stretching, squeezing, and random cut-and-paste techniques yielded similar distributions, specifically normal distribution.

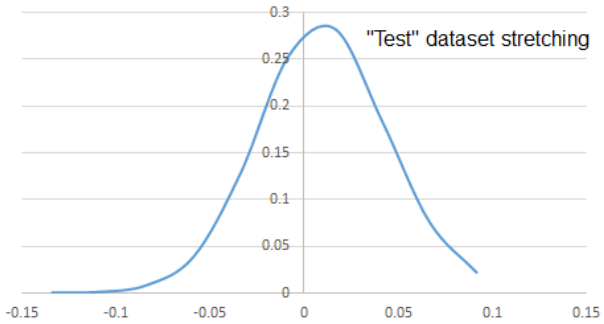

Fig. 28: After applying the stretching technique, the PMF of the augmented "test" dataset results in a normal distribution with parameters μ equal to 0.090909091 and σ equal to 0.011071521

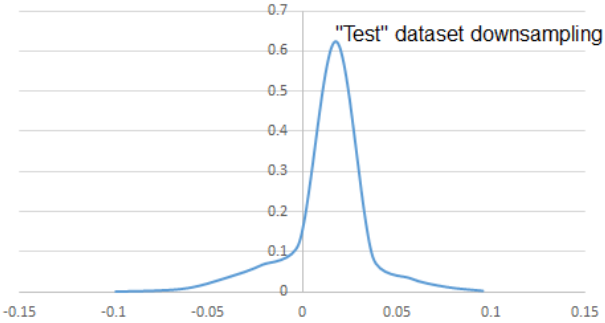

Fig. 29: After applying the squeezing technique (downsampling), the PMF of the augmented "test" dataset results in a normal distribution with parameters μ equal to 0.090909091 and σ equal to 0.032889586

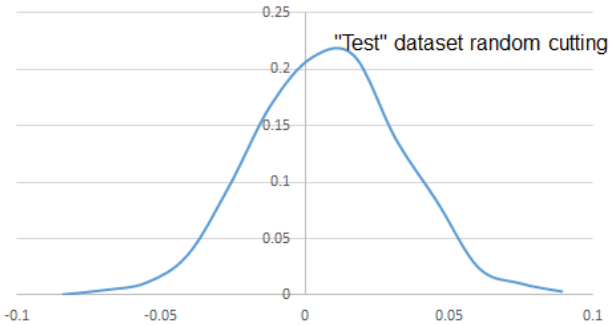

Fig. 30: After applying the random cut-and-paste technique, the PMF of the augmented "test" dataset results in a normal distribution with parameters μ equal to 0.076923077 and σ equal to 0.006554275

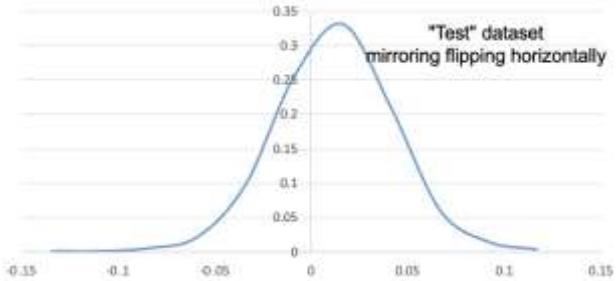

Fig. 31: After applying the mirroring flipping horizontally technique, the PMF of the augmented "test" dataset results in a normal distribution with parameters μ equal to 0.09091 and σ equal to 0.01415

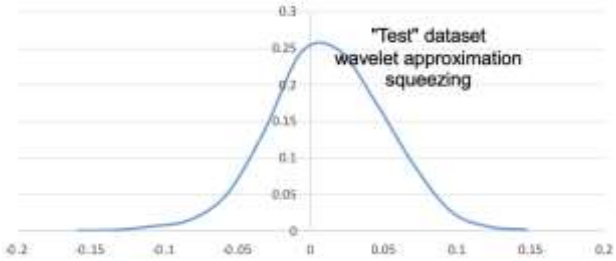

Fig. 32: After applying the squeezing technique (wavelet approximation coefficients), the PMF of the augmented "test" dataset results in a normal distribution with parameters: μ equal to -0.000743937 and σ equal to 0.001609512

7 Discussion on the Applications

The first application of detecting input-data properties, like long-range dependence, is a validation of the augmentation process. The second application of detecting input-data properties, like long-range dependence, is using proper machine learning tools to analyze the input data for extracting features, classification, or regression. Detecting the existence of long-range dependence in the input data is important to selecting or designing various structures of neural networks. If input data do not have long-range dependence, a regular neural network or a multiscale neural network structure can be sufficient for analyzing the data. If there is long-range dependence within the input data, designing a poly-scale neural network structure could be particularly useful. Consequently, a poly-scale analysis algorithm is recommended for designing the architecture of the poly-scale neural network structure, [19], [20], [21]. In contrast with the multiscale analysis, where there is no correlation between the outcomes at different scales, a poly-scale analysis measures input data at various scales, and its outcome requires all the scales to be used simultaneously [6]. Thus, the hidden feature in the input data could be extracted using the poly-scale neural network. The design of such a poly-scale neural network is addressed in [19].

Applications of data augmentation include all the deep learning and machine learning architectures [22], particularly in the *convolutional neural networks* (CNNs) models, whenever there are limited sizes of quality data and to improve the model's robustness and performance. It has been used in healthcare [23] and autonomous vehicles to expand the range of scenarios for self-driving cars and drones. It has also been used in *natural language processing* (NLP) to improve its performance by synonym augmentation, word embedding, character swap, and random insertion and deletion, [24]. Automatic speech recognition has also benefited significantly from data augmentation. Image processing and computer vision have also used data augmentation extensively. The traditional limitations of data augmentation include biases, [25]. Our paper identifies another serious limitation of data augmentation for data with long-term dependence. This insight highlights the need to refine augmentation practices in modern communications, enhancing the reliability and robustness of models used for network monitoring and security applications, [26], [27].

Moreover, this study shows some data augmentation methods and measures that can be used for other applications, such as particle filters. When using particle filters or sample-based methods in general, researchers need sampling. Instead of random sampling, one effective approach is employing a data augmentation method, such as wavelet approximation and detail coefficients. As a result, the area that has more information can be exaggerated. Then, the area of the input signal with more information has a larger amplitude. Conversely, areas without any information exhibit lower amplitude in the augmented data. Finally, the sampling process can be launched from the approximation coefficients area where its corresponding detail coefficients have high values.

One application of data augmentation techniques that preserves LRD is the effective management of big data in communications. By maintaining the fundamental properties of the original datasets, augmented data can enhance the performance of predictive models, anomaly detection systems, and traffic management strategies. This leads to more accurate analysis, improved decision-making, and greater operational efficiency in handling the complex, bursty, and highly dynamic nature of communications data.

Finally, the Hurst exponent and the VFDT can also be used to sample particle filters. In normal data, the boundaries of VFDT are different from the boundaries of anomalous data. Therefore, to extract more detailed information from the corresponding raw data, it is crucial to find the trajectory of the VFD, as this will help identify points that belong to anomalous boundaries, providing more insight into the underlying patterns. Therefore, more samples from this signal area with VFD and anomalous boundaries can be obtained.

8 Extension of the Work

In this paper, seven measures are considered to analyze the impact of data augmentation on datasets. To expand the analysis of the dataset, other measures can also be considered, such as the *discrete wavelet transform* (DWT), *principal component analysis* (PCA), the distance between a packet's arrival time with the adjacent packet, and the average of the distance between a packet's arrival time with the last five adjacent packets. These measures can be utilized to examine the time series signals of both pre and post-data augmentation.

For future work, exploring the combined use of these additional measures could provide deeper insights into the effectiveness of data augmentation techniques, particularly for datasets originating from stochastic environments like Internet traffic data.

9 Concluding Remarks

This study investigated the impact of five augmentation techniques on the long-range dependence of the DDoS ITD and the "test" dataset. It highlights how these techniques can disrupt the LRD of certain time series datasets by altering their PMF.

While none of the augmentation methods preserved the original data constraints of the DDoS ITD, as indicated by the changes in PMF, the augmented "test" dataset closely mirrored the original distribution, following a normal distribution and thus preserving the constraints. Consequently, the proposed augmentation techniques may be suitable for audio time series data but not for Internet time series data that originate from bursty operations.

This study also demonstrates how LRD can serve as a tool to assess the reliability and validity of the augmentation process. Moreover, this research illustrates the real-time applicability of using PMF to evaluate the suitability of data augmentation techniques. Furthermore, the PMF validation approach is adaptable to various time series datasets, offering significant convenience and advantageous in determining whether to expand input data for machine learning purposes.

Future work could investigate hybrid augmentation methods that integrate domain-specific constraints. This approach aims to enhance the robustness and applicability of augmented datasets in modern communications, particularly for tasks such as network anomaly detection and predictive traffic modeling.

Acknowledgement:
The National Science Foundation, the US Department of Homeland Security, and CAIDA Members provided funding for CAIDA's Internet Traces. Additionally, Sina Sedigh supported the University of Manitoba Speech Dataset.

Declaration of Generative AI and AI-assisted Technologies in the Writing Process

Maryam Ghanbari used ChatGPT and Grammarly to only edit this paper, enhancing its clarity and grammatical accuracy.
The authors reviewed and edited the content and take full responsibility for the content of this publication.

References:

[1] Maryam Ghanbari and Witold Kinsner, "Data augmentation methods and their effects on long-range dependence," in Proc. *of 20th IEEE International Conference on Cognitive Informatics and Cognitive Computing (ICCI*CC'20)*, Beijing, China, pp. 169–178, Sep 2020.

[2] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton, "Imagenet classification with deep convolutional neural networks," In *Advances in Neural Information Processing Systems 25 (NIPS 2012)*, Lake Tahoe, Nevada, USA, pp. 1097–1105, Dec 2012.

[3] Esben Jannik Bjerrum, Mads Glahder and Thomas Skov, "Data Augmentation of Spectral Data for Convolutional Neural Network (CNN) Based Deep Chemometrics," *arXiv.org (2017)*, Cornell University, NY, the USA, pp. 1–10, Oct 2017.

[4] The CAIDA UCSD "The CAIDA DDoS attack 2007 dataset," *caida.org*, 2015,

[Online]. https://www.caida.org/data/passive/ddos-20070804_dataset.xml (Accessed Date: September 26, 2024).

[5] Sen Xin Zhou, Jiang Hong Han, and Hao Tang, "A Trust Evaluation Model for Industrial Control Ethernet Network," *International Journal of Wireless and Microwave Technologies (IJWMT)*, vol. 1, no. 5, pp. 60–66, Oct 2011. https://doi.org/10.5815/ijwmt.2011.05.09.

[6] Witold Kinsner, *Fractal and Chaos Engineering: Monoscale, Multiscale and Polyscale Analyses*. Winnipeg, MB: OCO Research, Jan 2020, 1106 pages. ISBN: 978-0-9939347-1-1, pbk.

[7] David S. Moore, George P. McCabe, Bruce A. Craig, *Introduction to the Practice of Statistics*, 6th ed. W.H. Freeman and Company New York, 2009.

[8] John G. Proakis and Dimitris G. Manolakis, *Digital Signal Processing*, 4th ed. N.J.: Pearson Prentice Hall, 2007.

[9] Alan V. Oppenheim and Ronald W. Schafer, *Digital Signal Processing*, Prentice-Hall Inc. Englewood Cliffs, New Jersey, 1975.

[10] Natalia M. Markovich and Udo R. Krieger, "Statistical Analysis and Modeling of Peer-to-Peer Multimedia Traffic," D. Kouvatsos (Ed.): Next Generation Internet, LNCS 5233, Springer-Verlag, Berlin, Heidelberg, pp. 70–97, 2011.

[11] Esther Stroe-Kunold, Tetiana Stadnytsk, Joachim Werner, and Simone Braun, "Estimating long-range dependence in time series: An evaluation of estimators implemented in R," *Behav. Res. Methods*, vol. 41, no. 3, pp. 909–923, 2009. https://doi.org/10.3758/BRM.41.3.909.

[12] Jake M. Ferguson, Felipe Carvalho, Oscar Murillo-Garcia, Mark L. Taper, and Jose M. Ponciano, "An Updated Perspective on the Role of Environmental Autocorrelation in Animal Populations," Theoretical Ecology, vol. 9, no. 2, pp. 129–148, Aug 2015. https://doi.org/10.1007/s12080-015-0276-6.

[13] M. S Borella, S. Uludaq, G.B. Brewster, I. Sidhu, "Self-Similarity of Internet Packet Delay," in *Proc. of ICC'97 International Conference on Communications*, Montreal, Quebec, Canada, pp.513–517, 1997.

[14] Ingemar Kaj, *Stochastic Modeling in Broadband Communications Systems*. Philadelphia, PA, USA: Society for Industrial and Applied Mathematics, 2002.

[15] Rangaraj Rangayyan, *Biomedical Signal Analysis*, 1st ed., Wiley-IEEE Press, 2001.

[16] Madiha Jalil, Faran Awais Butt, and Ahmed Malik, "Short-time energy, magnitude, zero crossing rate and autocorrelation measurement for discriminating voiced and unvoiced segments of speech signals," in *Proc. of 2013 International Conference on Technological Advances in Electrical, Electronics and Computer Engineering (TAEECE 2013),* Konya, Turkey, pp. 208–212, May 2013.

[17] Ramalingam Shanmugam and Rajan Chattamvelli, *Statistics for Scientists and Engineers*, 1st ed. John Wiley & Sons, Incorporated, 2015, pp. 97–104.

[18] Amath 301, *Lecture: Polynomial Fits and Splines*, 2016, [Online]. https://www.youtube.com/watch?v=bFOTmSsDtAA (Accessed Date: September 26, 2024).

[19] Maryam Ghanbari and Witold Kinsner, "Detecting DDoS attacks using a policy gradient based deep reinforcement learning," in *Proc. of 21st IEEE International Conference on Cognitive Informatics and Cognitive Computing (ICCI*CC'21*), Banff, AB, Canada, pp. 158–165, Oct 2021.

[20] Ashraf A. Abu-Ein, Waleed Abdelkarim Abuain, Mohannad Q. Alhafnawi, and Obaida M. Al-Hazaimeh, "Security enhanced dynamic bandwidth allocation-based reinforcement learning," *WSEAS Transactions on Information Science and Applications*, vol. 22, no. 1, pp. 21–27, 2025. Available: https://wseas.com/journals/articles.php?id=9744.

[21] M. Sabrigiriraj and K. Manoharan, "Teaching machine learning and deep learning introduction: An innovative tutorial-based practical approach," *WSEAS Transactions on Advances in Engineering Educatio*n, vol. 21, no. 1, pp. 54–61, 2024. https://doi.org/10.37394/232010.2024.21.8.

[22] Connor Shorten and Taghi M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *J. of Big Data*, vol. 6, no. 60, Jul 2019, https://doi.org/10.1186/s40537-019-0197-0.

[23] Phillip Chlap, Hang Min, Nym Vandenberg, Jason Dowling, Lois Holloway, and Annette Haworth, "A review of medical image data augmentation techniques for deep learning applications," *J. of Medical Imaging and Radiation Oncology*, vol. 65, no. 5, pp. 545-563, Jun 2021. https://doi.org/10.1111/1754-9485.13261.

[24] Bohan Li, Yutai Hou, and Wanxiang Che, "Data augmentation approaches in natural language processing: A survey," *AI Open*, vol. 3, pp. 71-90, 2022, https://doi.org/10.1016/j.aiopen.2022.03.001.

[25] Alhassan Mumuni and Fuseini Mumuni, "Data augmentation: A comprehensive survey of modern approaches," *Array*, vol. 16, no. 100258, Dec 2022, https://doi.org/10.1016/j.array.2022.100258.

[26] Elie El Ahmar, Ali Rachini, and Hani Attar, "Cybersecurity enhancement in IoT wireless sensor networks using machine learning," *WSEAS Transactions on Information Science and Applications*, vol. 21, no. 1, pp. 480–487, 2024. https://doi.org/10.37394/23209.2024.21.43.

[27] Nabeel Refat Al-Milli and Yazan Alaya Al-Khassawneh, "Intrusion Detection System using CNNs and GANs," *WSEAS Transactions on Computer Research*, vol. 12, no. 1, pp. 281–290, 2024, https://doi.org/10.37394/232018.2024.12.27.

Contribution of Individual Authors to the Creation of a Scientific Article (Ghostwriting Policy)

Conceptualization, M.G., and W.K.; Data Curation, M.G.; Investigation, M.G; Writing—original draft preparation, M.G.; Writing—review and editing, M.G., W.K. and N.S.; Supervision, W.K.; Project Administration, W.K., and N.S.; funding acquisition, W.K., N.S.; All authors have read and agreed to the published version of the manuscript.

Sources of Funding for Research Presented in a Scientific Article or Scientific Article Itself

This research was partially funded by the Natural Sciences and Engineering Research Council of Canada (NSERC). Grant number: RGPIN-2018-05352.

Conflict of Interest

The authors declare no conflicts of interest.

APPENDIX

Table 1. The statistical data of the DDoS ITD and the results of eight mono-scale, multi-scale, and poly-scale measures both before and after data augmentations for the initial stationary frame of trajectories

	Raw Data	**Flipping Horizontally (Mirroring)**	**Stretching**	**Squeeze Downsampling**	**Squeeze Wavelet transform**	**Random Cut-and-Paste**	**Adding Noise**
Mean	0.0029	0.0029	0.0027	0.0014	0.0041	0.0030	0.0029
Median	0.0020	0.0020	0.0023	0	0.0038	0.0021	0.0029
Mode	5.0000e–06	5.0000e–06	6.0000e–06	0	–0.0014	5.0000e–06	–0.0340
Variance	7.7418e–06	7.7418e–06	4.4139e–06	5.9432 e–06	6.4685e–06	8.6087e–06	1.1161e–04
Standard deviation	0.0028	0.0028	0.0021	0.0024	0.0025	0.0029	0.0106
Skewness	1.2928	1.2928	1.0344	2.1488	0.4927	1.2740	0.0804
Kurtosis	1.2582	1.2582	0.9338	4.6103	–0.2094	1.1205	3.8383e–04
Autocorrelation (lag 0)	–3.4694e–18	–3.4694e–18	0	5.2042e–18	3.0095e–07	8.5698e–07	–1.0536e–05
VFD	1.9931	1.9931	1.9836	1.9880	1.9891	1.9944	1.9864
Hurst exponent	0.0069	0.0069	0.0164	0.0120	0.0109	0.0056	0.0136
Slope (β)	1.0138	1.0138	1.0327	1.0240	1.0218	1.0111	1.0272
Zero crossing	1250	1250	1008	598	3154	1228	4129
Turns count	5349	5349	5516	2846	6017	5240	5472

Table 2. The statistical information of the word “test” dataset and the results of eight mono-scale, multi-scale, and poly-scale measures both before and after data augmentations for the initial stationary frame of trajectories

	Raw Data	**Flipping Horizontally (Mirroring)**	**Stretching**	**Squeeze Downsampling**	**Squeeze Wavelet transform**	**Random Cut-and-Paste**	**Adding Noise**
Mean	–0.0051	–0.0052	–0.0060	–0.0025	–0.0066	–0.0055	–0.0054
Median	–0.0055	–0.0055	–0.0050	0	–0.0078	–0.0013	–0.0073
Mode	–0.0429	–0.0429	–0.0050	0	–0.1583	–0.0119	–0.1361
Variance	0.0015	0.0015	0.0017	7.3658e–04	0.0026	0.0013	0.0016
Standard deviation	0.0383	0.382	0.0410	0.0271	0.0508	0.0358	0.0397
Skewness	0.0917	0.0890	0.0834	0.0068	0.0525	0.0151	0.1513
Kurtosis	0.3912	0.4106	0.2173	2.7957	0.6188	0.2191	0.4643
Autocorrelation (lag 0)	0.0010	0.0010	–0.0003	0	–0.0024	–4.0610e–05	0.0025
VFD	1.9412	1.9412	1.9273	1.9867	1.9835	1.9007	1.9767
Hurst exponent	0.0588	0.0588	0.0727	0.0133	0.165	0.0993	0.0233
Slope (β)	1.1177	1.1117	1.1455	1.0267	1.0330	1.1987	1.0466
Zero crossing	145	145	139	123	250	114	159
Turns count	192	192	418	150	290	171	226