Comparative Analysis between Symmlets and Haar functions for the Study of De-noising with Application to Financial Time Series Data

Amel Abdoullah Ahmed Dghais
Faculty of accounting University Garin
(E-mail: Dr.amel.usm@gmial .com).

Abstract— since the day wavelet came into the picture, the field of wavelets has been growing on both the theoretical and applications fronts. Therefore, in this paper, the core objective is to study the differences in order to obtain less noise data between the two wavelets functions. The wavelet functions are Haar and Symmlets4 based on discrete wavelet transform (DWT) using autocorrelation function (ACF) and peck signal to noise ratio (PSNR). Therefore, to achieve this, data from Dow Jones index (DJIA30) of US stock market are being utilized. The data consists of 2048 daily closing prices starting from December 17, 2004 until October 23, 2012. Results indicate that Symmlets4 function produced less noise series as compared to Haar function

Keywords—*DWT, PSNR, ACF, T-test*

الملخص:

منذ اليوم الذي ظهر فيه تحويل الموجات في الصورة، كان مجال الموجات ينمو على الجهتين النظرية والتطبيقات. لذلك، في هذا البحث، الهدف الأساسي هو دراسة الاختلافات بين الدالة **Sym4** و الدالة **Haar** بواسطة تحويل الموجات المنفصل (**DWT**) لأجل تنقية السلسلة الزمنية من الضوضاء والحصول على سلسلة زمنية ذات ضوضاء أقل , حيث تم استخدام دالة الارتباط الذاتي (**ACF**) و نسبة الذروة إلى الضوضاء متوسط مربعات الخطأ التربيعي (**PSNR**) للمقارنة بين الدالتين . حيث استخدمت الدراسة بيانات السلاسل الزمنية من سوق الاسهم الامريكي داو جونز (**DJIA30**) , حيث تحتوي البيانات على 2048 سعر إغلاق يومي تبدأ من 17 ديسمبر 2004 وحتى 23 أكتوبر 2012. تبين النتائج بناء على (**ACF**) و(**PSNR**) إلى أن الدالة **Sym4** أنتجت سلسلة زمنية ذات ضوضاء أقل مقارنة بالدالة **Haar**.

I. INTRODUCTION

Wavelet transform has become an effective and fast-growing technique for representing real-life non-stationary signals with great efficiency. Wavelet transform has application in diverse areas, such as physics, engineering, signal processing, applied mathematics, and statistics. Given that financial time series is governed by nonlinear and non-stationary behaviour, capturing the prevalent features of their fluctuation has become a challenge. Therefore, wavelet transform has become an alternative tool for studying the behaviour of financial datasets. Generally speaking, the wavelet transform is a tool that partitions data into different frequency components and then studies each component with a resolution matched to its scale (Daubechies, 1992). Therefore, it can provide economical and informative mathematical representation of many objects of interest (Abramovich et al., 2000). Moreover, wavelet transform has been dominating the standard time series econometric tools, such as Fourier transform, because the latter considers only time or frequency components, whereas wavelets allow the study of both components simultaneously. Consequently, wavelets can reveal the interactions between the time and frequency components of time series data, where econometric tools fail to do. In the study about financial datasets (Razak and Ismail 2010), the authors have compared four wavelets functions, namely Haar, Daubechies, Symmlets, and Coiflets, of DWT and MODWT by applying them to Malaysian stock prices. Their results exposed that the Daubechies and Symmlets are the best functions with

DWT whereas Daubechies is better than Symmlets with MODWT .In this regard, (Malik and Verma 2012) made comparisons between the Haar and Daubechies of DWT and between the DWT and discrete cosine transform (DCT), revealing that DWT is the best transform and Daubechies produces better results.

Seven true-color images were used to identify which of the eleven different wavelet filters illustrate the effects on the images (Abbas 2012). Based on the peak signal to noise ratio (PSNR), the results suggested that the Daubechies family is better in compressing images. (Chavan and Mastorakis2010) examined wavelet transform performance by Haar and Daubechies in speech denoising data. They discovered that higher order Daubechies wavelet is suitable for speech denoising. (Singh et al 2011) which compared three different wavelet families for image compression, based on peak signal-to-noise ratio, they concluded that the biorthogonal wavelet is the best function among all the families for low pixel size image. However, for high pixel size image Coiflet is better suited. In case of medium size images, both biorthogonal and Daubechies provides better results. (Bolzan et al 2008) explored the performance of two wavelet functions namely Daubechies and Haar in extracting the coherent structures from solar wind velocity time series. They found that both wavelet functions are able to extract coherent structures, however, the coherent time series showed that the Daubechies wavelet function was able to extract more coherent structures than the Haar wavelet. On the other hand, (Zaheer et al 2018) presented the introducing fundamental ideas connected with wavelets and data mining and a survey of applications of wavelets to various aspects data mining. (Reginald et al 2019) used Morlet wavelets to discover the morphology of a time series cyclical components and the unsupervised data mining of financial time series in order to discover hidden motifs within the data. In addition, the results proposed the implementation of the "Bolman Time Series Power Comparison" algorithm extracted the pertinent time series motifs from the underlying dataset.

Our main purpose will be to investigate the difference between Haar and Symmlets4 to product less noise data of DWT. It is revealed thru from the results that the Haar is the lowest smoothing as compared to Symmlets4.

The remainder of this paper is organized as follows. Section II briefly discusses the methodology. Section III describes the data, the empirical results and discussion. Finally, section IV concludes this paper.

II. METHODOLOGY

A. Wavelets

The wavelets have two types, the father wavelets ϕ (Heil and Walnut 1989) and the mother wavelets $\psi$where father wavelet ϕintegrates to one and mother waveletψintegrates to zero (Mallat 1989). That is

$$\int \phi(t)\,dt = 1 \quad \text{and} \quad \int \psi(t)\,dt = 0 \qquad (1)$$

The mother wavelets are useful in describing the detail and high-frequency components while the father wavelets are good at representing the smooth and low-frequency parts of signal.

Wavelets are derived using a special two-scale dilation equation. Father wavelet$\phi(t)$ and mother $\psi(t)$ are defined as

$$\phi(t) = \sqrt{2}\sum \ell_\kappa \phi(2t-k) \qquad (2)$$

$$\psi(t) = \sqrt{2}\sum h_\kappa \phi(2t-k) \qquad (3)$$

Where$\ell_\kappa$and$h_\kappa$defined in equation (4) and (5) respectively, are low-pass and high-pass filter coefficients used to pass the original signal.

$$\ell\kappa = \frac{1}{\sqrt{2}}\int \phi(t)\phi(2t-k)dt \quad (4)$$

$$h\kappa = \frac{1}{\sqrt{2}}\int \psi(t)\phi(2t-k)dt \quad (5)$$

The wavelet series approximation to a signal $X(t)$ is defined by:

$$X(t) = \sum_k S_{J,\kappa}\phi_{J,\kappa}(t) + \sum_k d_{J,\kappa}\,\psi_{J,\kappa}(t) + \sum_k d_{J-1,\kappa}\,\psi_{J-1,\kappa}(t) + \ldots + \sum_k d_{1,\kappa}\,\psi_{1,\kappa}(t) \quad (6)$$

Where k ranges from 1 to the number of coefficients in the specified components (or crystals) and J is the number of multi resolution levels (or scales). The coefficients $S_{J,\kappa}$,$d_{J,\kappa},\ldots,d_{1,\kappa}$ are wavelet transform coefficients given by the projections

$$S_{J,\kappa} = \int \phi_{J,\kappa}(t)\,X(t)dt \quad (7)$$

$$d_{j,\kappa} = \int \psi_{j,\kappa}(t)\,X(t)dt, \quad j = 1, 2, .J \quad (8)$$

The magnitude of these coefficients gives a measure of the contribution of the corresponding wavelet function to the total signal. The basic functions $\psi_{j,\kappa}$ and $\phi_{j,\kappa}$, $j = 1,2,.J$ are the approximating wavelet functions generated as scaled and translated versions of ϕ and ψ with scale factor 2^j and translation parameter $2^j k$ respectively, defined as:

$$\phi_{j,\kappa} = 2^{\frac{-j}{2}}\phi\left(2^{-j}\,t-\kappa\right) = 2^{\frac{-j}{2}}\phi\left(\frac{t-2^j\kappa}{2^j}\right) \quad (9)$$

$$\psi_{j,\kappa}(t) = 2^{\frac{-j}{2}}\psi\left(2^{-j}\,t-\kappa\right) = 2^{\frac{-j}{2}}\psi\left(\frac{t-2^j\kappa}{2^j}\right),\ j = 1,2,..J \quad (10)$$

Translation parameter $2^j\kappa$ is matched to the scale parameter 2^j in a way that as the function $\phi_{j,\kappa}$ and $\psi_{j,\kappa}(t)$ get wider, their translation steps are correspondingly larger.

The aim of discrete wavelet transform is to decompose the discrete time signal to basic functions called the wavelets, to give us a good analytic view of the analyzed signal. DWT is used to calculate the coefficients of approximation in equation (6) for a discrete signal of final extent $f_1, f_2, \ldots, f_n$.. It maps the vector $\mathbf{f} = (f_1, f_2, \ldots, f_n)'$ to a vector of n wavelet coefficients

$\boldsymbol{\omega} = (\omega_1, \omega_2, \ldots, \omega_n)'$ that contains both the smooth coefficient $S_{J,\kappa}$ and the detail coefficients $d_{j,\kappa}$, $j=1,2,..,J$. $S_{J,\kappa}$ representing the underlying smooth behavior of the signal at the coarse scale 2^j.On the other hand $d_{j,\kappa}$ describe the coarse scale deviations and $d_{J-1,\kappa},\ldots,d_{1,\kappa}$ provide progressively finer scale deviations. When the length of the signal n is divisible by 2^j, there are n/2 coefficients $d_{1,\kappa}$ at the first scale $2^1 = 2$. At the next finest scale $2^2 = 4$, there are $n/4$ $d_{2,\kappa}$ coefficients. Similarly, at the coarsest scale, there are $n/2^j$ $d_{j,\kappa}$ *and* $n/2^J$ $S_{J,\kappa}$ and coefficients. Altogether there are total of n coefficients:

$N = n/2 + n/4 + \ldots + n/2^{J-1} + n/2^J$.

B. Unit Root Test

All random processes are consists of random variables, each having its own vary point in time. Hence such processes have all the properties of random variables, such as correlation, variances , mean, etc. so , it is important in time series financial analysis to test whether theses statistical properties hold true for the entire random process. To facilitate this, the concept of stationary processes has been ordered. The random process

is called stationary where all its statistical properties don't change over time. However, process whose statistical properties vary with time is called non-stationary processes. Two types of unit root the Augmented Dickey Fuller (ADF) (Dickey and Fuller 1981) and Phillips Perron (PP) (Phillip and Perron 1988) tests are used in this paper to test for stationarity. These tests are defined by:

$$\Delta y_t = (P_a - 1)y_{t-1} + \mu_t \qquad (11)$$

$$\Delta y_t = a + \beta y_t + \varepsilon_t \qquad (12)$$

C. Autocorrelation Function (ACF)

ACF is a mathematical tool that is usually used for analyzing functions or series of values, for example time series signals and to measure the correlation between the signals.

ACF is a correlation coefficient. Nevertheless, instead of correlating between two different variables, the correlation is between two values of the same variable at times y_i and y_{t+k} . ACF is defined as:

$$p_k = \frac{E\left((y_t-\mu)(y_{t+k}-\mu)\right)}{\sqrt{E\left((y_t-\mu)^2 E(y_{t+k}-\mu)^2\right)}} = \frac{Cov(y_t, y_{t+k})}{Var(y_t)} = \frac{\gamma_k}{\gamma_0}, \quad k=0,1,2,.. \quad (13)$$

Note that by definition $p_0 = 1$.

Furthermore, the ACF is used to detect the stationary or non-stationary data, by observing the behaviour of the autocorrelation function. A strong and slowly dying ACF clearly suggests deviations from stationarity. Due to cutting off or tailing off near zero after a few lags the ACF is very persistent , meaning that it decays very slowly and exhibits sample autocorrelations that are still rather large even at long lags. This behaviour is characteristic of non-stationary time series. However, if the ACF values approaching to 1 it indicate the series is near to stationary and less noisy (Montgomery and Kulahci 2008). From a time series of finite length, the autocorrelation function is estimated as:

$$r_k = \frac{\sum_{t=1}^{T=k}(y_t - \bar{y})(y_{t+k} - \bar{y})}{\sum_{t=1}^{T=k}(y_t - \bar{y})^2}, \quad k = 0,1,2,...,K \qquad (21)$$

D. PSNR

Peak Signal to Noise Ratio (PSNR) represents a measure of the peak error and is expressed in decibels. It is defined by:

$$PSNR = 10\log_{10}\frac{(255^2)}{MSE}$$

Where $MSE = \frac{1}{mn}\sum_{i=0}^{m=1}\sum_{j=0}^{n=1}\left[x(i,j) - x_c(i,j)\right]^2$

Where $x(i,j)$ and $x_c(i,j)$ are the signal data and its corresponding data compression. The higher the value of the PSNR, the better the quality of the compressed or reconstructed signal (Suchitra Oinam et al 2013) .

III. EMPIRICAL RESULTS AND DISCUSSION

In this section the first given the description the data then the paper apply the Unit Root test to check if the data is stationary or not, thereafter the work chose two wavelet families toapply DWT with closing price of Dow Jones index (DJIA30)

A. Data discretion

The data used in the following analysis consists of the daily prices of the Dow Jones Index (DJIA30), for the period December 2004 to October 2012. The DJIA30 was selected primarily because of its long history and global significance on capital markets. This index is regarded as a certain mood indicator on capital markets. It is the representative of average price development in international markets and its Industrial average tracks the prices of 30 blue-chip stocks which represent 27 percent of the total US stock market value.

It should be noted that daily data is used in this paper because it does not lose any information regarding behaviour, and the daily prices on a given calendar day may represent prices realized over different time intervals depending on holidays and trading day schedules.

B. Stationarity Tests

Figure (1) illustrates the stock market index behaviour over time. It can be seen that the data is not stationary and it appears that from the end of 2004 until the end of 2005 the prices are quite stable. However, they start to increase until they hit an intra-day peak of 14,198.10 in October 9, 2007. Then the prices begin to decline starting from the beginning 2008 until they reach an intra-day low of 6,469.95 on March 9, 2009 as a result of the Global Financial Crisis of 2008–2009. After that the prices start to increase again until the middle of 2011 before decreasing in September 2011 because of another financial crisis. Figure (2) illustrates the index are quite stationary after first differences without any trend and the series fluctuates around zero.

The results of two tests (ADF and PP) using unit root test are given in Table (I). The P-value > 0.05 indicates that the data has unit root and therefore is non-stationary at level. However, it is stationary after the first different.

Fig: 1 Original Daily Price

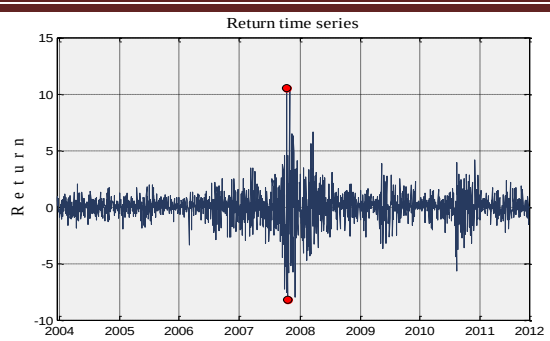


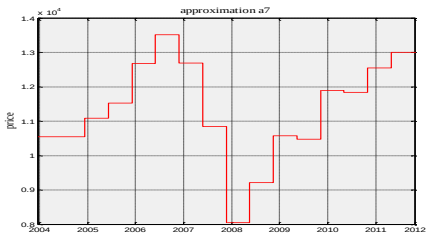
Fig 2: Return Daily Price

TABLE I
Unit Root Test Results of Original Series of Stock Market Indices

(ADF)		(PP)	
p-value at level	p-value at first difference	p-value at level	p-value at first difference
0.7896	0.00000*	0.8210	0.00000*

C. Wavelet Decomposition

In this section, the fingers 3 and 4 presents the (DWT) by two functions (Haar) and Symmlets4 (Sym4) for original daily price of DJIA30,the number of scales j=7 produces seven vectors of wavelet filter coefficients d1,d2,d3,d4,d5,d6,d7 and one vector of scaling coefficients a7, the result shows that, the first two/three high frequency components (d1,d2,d3) explain the higher part of the fluctuation of the series. In addition, the every a7 of the function is very close to the original series. However, there is no much different among each other. However, from the figures cannot indicate the best function for decomposition the data which will give less noise data.



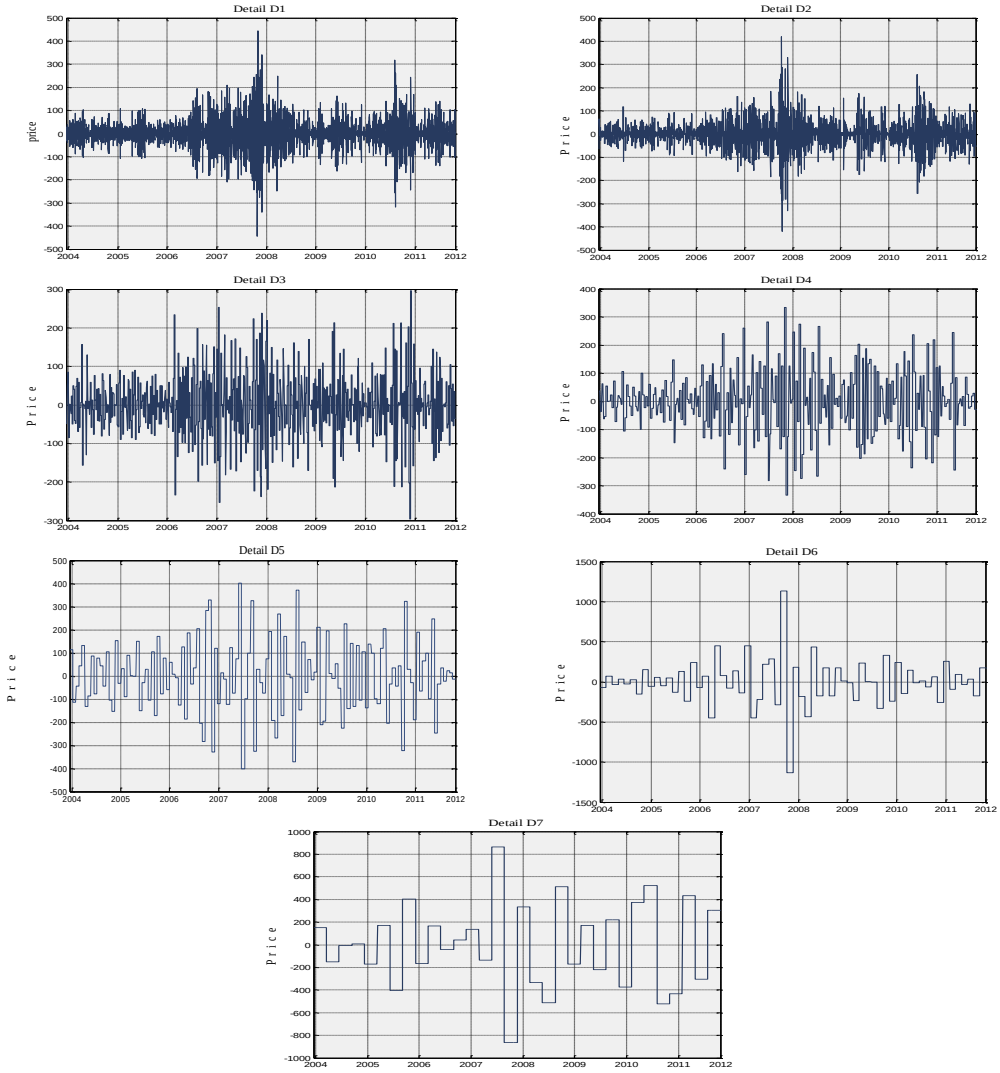
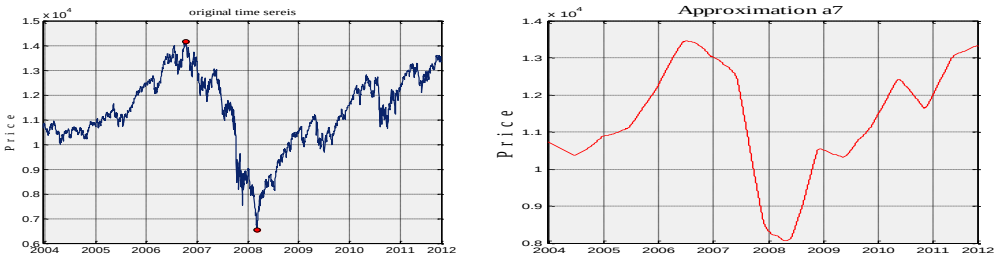


Fig 3: Analysis by DWT (Haar) for Dow stock market



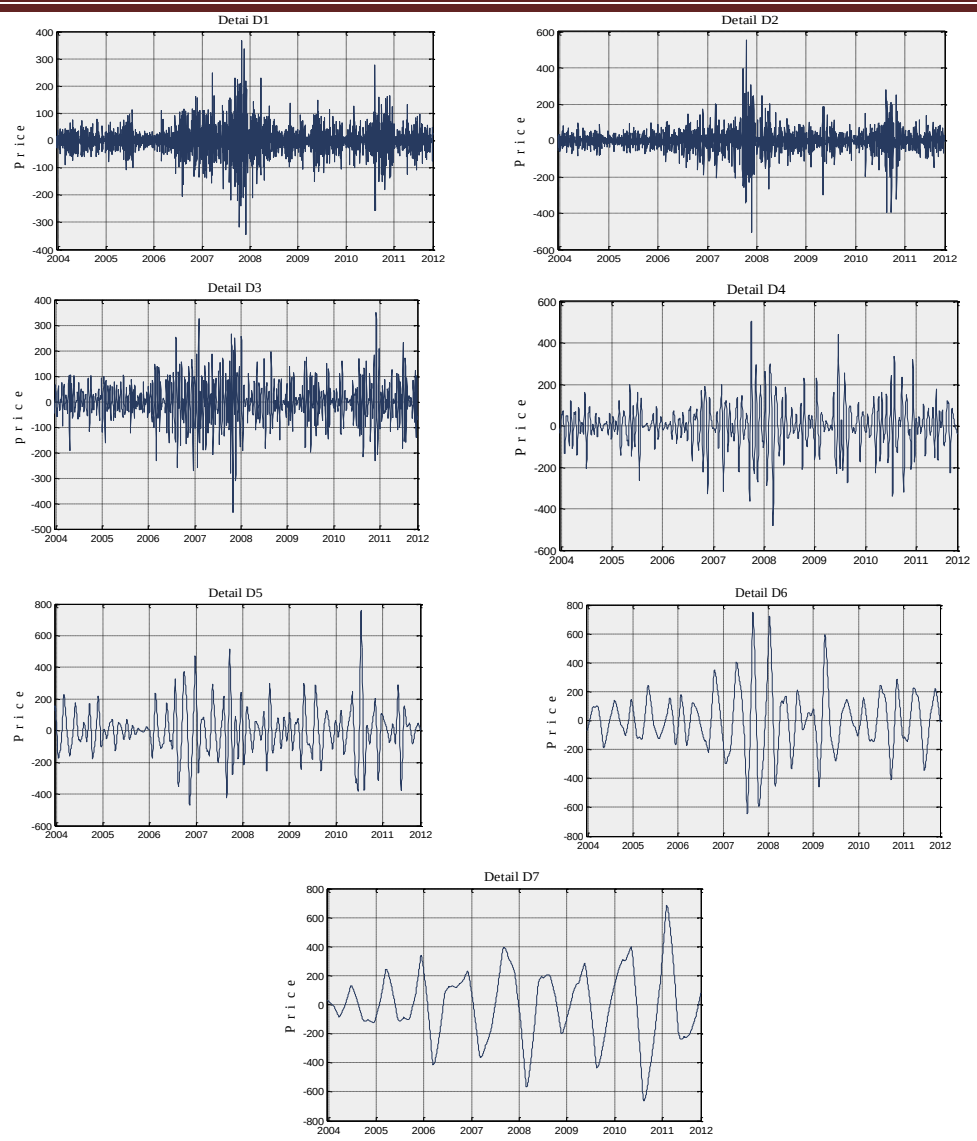


Fig 4: Analysis by DWT (sym4) for Dow stock market

D. Autocorrelation (ACF)

As from the figures, we cannot see clearly which is the best function, thus the paper applied the ACF as the best method to analyze the fluctuation in data. Consequently, if the autocorrelation is high (low) then the stationarity will be high (low) as the value approaching one (Booth and Koutmos 1998).

TABLE II

Autocorrelation values for data

lags	Haar	Sym4
0	1	0.999
1	0.997	0.999
2	0.994	0.998
3	0.991	0.997
4	0.988	0.997
5	0.985	0.996

6	0.982	0.995
7	0.979	0.994
8	0.976	0.993
9	0.973	0.992
10	0.971	0.991
11	0.968	0.99
12	0.965	0.988
13	0.962	0.987
14	0.959	0.986
15	0.956	0.984
16	0.953	0.983
17	0.95	0.999
18	0.947	0.982
19	0.944	0.98
20	0.941	0.978

Table II illustrates that after decomposition by sym4 and Haar based on DWT, the lowest wavelet is Haar compared with sym4. Since the ACF of Haar is the lowest function for different lags of the data. However, the difference between ACF values is very small; the T-test is used to ascertain that which of the two functions is better to be used for decomposition data in order to obtain less noisy data. The T-test was performed using two samples of data as follows:

H_0: $\mu_{Haar} = \mu_{sym4}$ $H_{1:}$ $\mu_{Haar} \neq \mu_{sym4}$

From T-test the P-value =0.000 and alpha=0.05. Since the P-value less than alpha value, then the differences are significant. This means that the functions are not the same in given less noise data

TABLE III
Mean of ACF for DJIA stock market using DWT

Wavelet functions	N	mean
Haar	21	0.926000
Sym4	21	0.978348

Table III the results shows that the Haar is the low mean, which in turn indicates that the sym4 function is better than Haar function to obtain less noisy data, which is importing for data mining.

E. **Peak Signal to Noise Ratio (PSNR*)***

In table IV the results shows that, the PSNR value of Haar less than PSNR value of sym4, which conclude that, the sym4 function is better than Haar function to obtain less noisy data.

TABLE IV
PSNR for DJIA stock market using DWT

Function	PSNR
Haar	9.7319
Sym4	11.8257

IV. CONCLUSION

Wavelet techniques have many advantages and there already exists numerous successful applications in Data Mining. It goes without saying that wavelet approaches will be of growing importance in Data Mining. Therefore, this paper aims to a comparison between two wavelet functions are Haar and Symmlets4 based on discrete wavelet transform (DWT), using original data of US stock market namely DJA30 to study the difference less noisy between two wavelet functions, The results based on ACF and PSNR revealed that, the sym4 gives less noisy data than Haar function in original dataset. That because, the Haar is the least smooth wavelet because it has only one vanishing moment. This result is important in Data Mining field.

References

[1] Daubechies, I.: 'Ten Lectures on Wavelets Capital City Press', Montpelier, Vermont, 1992.

[2] Abramovich, F., Bailey, T.C., and Sapatinas, T.: 'Wavelet analysis and its statistical applications', Journal of the Royal Statistical Society: Series D (The Statistician), 2000, 49, (1), pp. 1-29

[3] Razak, A., Aripin, R., and Ismail, M.T.: 'Denoising Malaysian time series data: A comparison using discrete and stationary wavelet transforms', in Editor (Ed.)^(Eds.): 'Book Denoising Malaysian time series data: A comparison using discrete and stationary wavelet transforms' (2010 International Conference on Science and Social Research (CSSR) IEEE, 2010, edn.), pp. 412-415.

[4] Malik, M.S., and Verma, M.V.: 'Comparative analysis of DCT, Haar and Daubechies Wavelet for Image Compression', International Journal of Applied Engineering Research, 2012, 7, ((11)), pp. 0973-4562

[5] H.K.Abbas: 'comparison of wavelet transform filters using image compression', Ibu Al-Haitham Journal of pure and applied science, 2012, 25, ((1))

[6] Chavan, M.S., and Mastorakis, N.: 'Studies on Implementation of Haar and daubechies Wavelet for Denoising of Speech Signal', International Journal of Circuits, Systems and Signal Processing, 2010, 4, ((3))

[7] Singh, P., Singh, P., and Sharma, R.K.: 'JPEG image compression based on biorthogonal, coiflets and daubechies wavelet families', International Journal of Computer Applications, 2011, 13, (1), pp. 1-7

[8] Bolzan, M., Guarnieri, F., and Vieira, P.C.: 'Comparisons between two wavelet functions in extracting coherent structures from solar wind time series', Brazilian Journal of Physics, 2009, 39, ((1)), pp. 12-17

[9] Raina, Z.A.P.: 'A Study on Applications of Wavelets to Data Mining', International Journal of Applied Engineering Research, 2018, 13, (0973-4562), pp. 10886-10896.

[10] Boucher, R.B.T.: 'Data Mining using MorletWavelets for Financial Time Series', In Proceedings of the 8th International Conference on Data Science, Technology and Applications, 2019, pp. 74-83

[11] Heil, C.E., and Walnut, D.F.: 'Continuous and discrete wavelet transforms', Society for Industrial and applied mathematics review 1989, 31, ((4)), pp. 628-666

[12] Mallat, S.G.: 'Multiresolution approximations and wavelet orthonormal bases of L2(R)', Transactions of the American Mathematical Society, 1989, 315, (1), pp. 69-87

[13] Dickey, D.A., and Fuller, W.A.: 'Likelihood ratio statistics for autoregressive time series with a unit root', Econometrica: Journal of the Econometric Society, 1981, 49, ((4)), pp. 1057-1072

[14] Phillips, P.C., and Perron, P.: 'Testing for a unit root in time series regression', Biometrika, 1988, 75, ((2)), pp. 335

[15] Montgomery, D.C., Jennings, C.L., and Kulahci, M.: 'Introduction to time series analysis and forecasting' (Wiley-interscience, 2008.

[16] Oinam, S., and HK P, P.S.: 'Compression of time series signal using wavelet decomposition, wavelet packet and decimated discrete wavelet compression transforms techniques and their comparison', Int J Adv Res Comput Commun Eng, 2013, 2, pp. 1540-1544

[17] Booth, G.G., and Koutmos, G.: 'Volatility and autocorrelation in major European stock markets', The European Journal of Finance, 1998, 4, (1), pp. 61-74