Wavelet Transform Data Analysis Tool

Orçun Çelik, Altuğ Numan Yıldız
Bilgisayar Mühendisliği Bölümü
Yıldız Teknik Üniversitesi, 34220 Istanbul, Türkiye
{l1115053, l1115064}@std.yildiz.edu.tr

***Özetçe* —Bu çalışmada dalgacık dönüşümü kullanılarak veri analiz aracı geliştirilmiştir. Dalgacık dönüşümü, frekansın zamanla değiştiği sinyallerde, frekansların hangi zaman aralıklarında olduğunu bulmak için kullanılır. Proje kapsamında kullanılan GTZAN veri seti, ayrık ve sürekli dalgacık dönüşümüne tâbi tutulduktan sonra farklı istatistiksel fonksiyonlar ile analiz edilmiştir. Analiz verileri üzerinden farklı makine öğrenmesi algoritmalarının başarım, hassasiyet ve duyarlılıkları test edilmiştir. En doğru sonuç (%67.25), Random Forest algoritması ile sym dalgacık ailesinde, ayrık dalgacık dönüşümü ile 3. seviyede elde edilmiştir.**

Anahtar Kelimeler—Dalgacık Dönüşümü, Sürekli Dalgacık Dönüşümü, Ayrık Dalgacık Dönüşümü, İstatistiksel Fonksiyonlar, Makine Öğrenmesi, Naif Bayes, Rassal Orman, Destek Vektör Makinesi

***Abstract*—In this study, the data analysis tool has been developed using wavelet transform. Wavelet transform is used to find the time intervals of frequencies in signals where the frequency changes with time. Within the scope of the project GTZAN data set was analyzed with different statistical functions of the discrete and continuous wavelet transform of the data set. Performance, precision and sensitivity of different machine learning algorithms were tested through analysis. The highest accuracy (67.25%) was achieved at the 3rd level of the discrete wavelet transform in sym wavelet family with Random Forest algorithm.**

Keywords—Wavelet Transform, Continuous Wavelet Transform, Discrete Wavelet Transform, Statistical Functions, Machine Learning, Naive Bayes, Random Forest, Support Vector Machine

I. Introduction

The analysis and processing of signals has an important place in human life. The Fourier transform reveals the frequency components of the signals, but not when the frequencies occur. STFT, another version of the Fourier transform, may have time intervals of frequency components, but frequency/time resolution problems arise due to the fixed window size. Wavelet transform provides a solution to the frequency/time resolution problem thanks to its changing window size. [1] The wavelet transform applications are listed below.

- 1D signal processing
- 2D signal processing
- Data Compression
- Medical Imaging

In this study, to compare performance results, all wavelet functions have been applied on the GTZAN dataset. The processed data have been saved to a database and by using those data, different classification algorithms were implemented. The classification algorithm that gives the most accurate result was determined. In the second section, wavelet functions which can be used for processing and extracting features of a signal such as standart deviation, kurtosis, skewness, mean etc. has been described. The classification algorithms have been explained in the third section. Experimental results has been summarized and discussed in the fourth section and the last section consist of conclusion.

II. Wavelet Types and Families

Wavelet transform is divided into two parts as discrete and continuous wavelet transform.

A. Discrete Wavelet Transform

Continuous wavelet transformation requires computation for every x shift value and every y scale value. Therefore, a lot of unnecessary calculations are required for continuous wavelet transform. Continuous wavelet transformation is difficult to implement in practice.

By decomposing x and y values, a transition from continuous wavelet transform to discrete wavelet transform is provided. In general, the parameters of the scale and translation are chosen as powers of 2.

Discrete wavelet transform can be explained by multiple resolution analysis. Multiple resolution analysis is an approach that calculates the approximation of the sign with orthogonal projections in different spaces. In this method, the L space is divided into a series of subspaces and each subspace is associated with a different scale.

$$W(m,n) = 2^{-m/2} \int f(t)\varphi\left(2^{-m}t - n\right)dt \tag{1}$$

The equation above gives the discrete wavelet transformation of a signal where m is scale, n is time, φ is the wavelet function and f(t) is the given input signal.

1) Haar: Haar is a wavelet family or array of "square shaped" functions rescaled on its basis. Haar explained the most important property of the wavelet function as it is limited and has tight support. However, the Haar wavelet function does not have a continuous derivative as db1. The same wavelet as Daubechies db1 is defined by Haar. [2]

$$\Psi(t) = \begin{cases} 1 & ,0 \le t < 1/2 \\ -1 & ,1/2 \le t < 1 \\ 0 & ,\ other \end{cases} \tag{2}$$

2) Daubechies: Daubechies wavelet is the most important wavelet type among wavelet transformations. They do not have clear statements. They are not symmetrical and the analysis is orthogonal.

$$\begin{aligned}\emptyset(t) &= \sqrt{2}\Sigma_k h_k \emptyset(2t-k)\\ \Psi(t) &= \sqrt{2}\Sigma_k g_k \emptyset(2t-k)\end{aligned} \tag{3}$$

3) Biorthogonal: They do not have clear statements. They are symmetrical and the analysis is biortogonal. Synthesis is fully possible with biortogonal wavelets, but not with orthogonal wavelets.

4) Coiflets: They are close to symmetry and analysis is orthogonal. Coiflet wavelet length is 6N-1, including N degrees. Both continuous and discrete wavelet transform can be done.

$$\int_{-\infty}^{+\infty} \emptyset(t)dt = 1 \text{ ve } \int_{-\infty}^{+\infty} t^k \emptyset(t)dt = 0 \quad , \quad 1 < k < p \tag{4}$$

5) Symlets: They are similar to symmetry and are orthogonal in analysis. The symmetric wavelet length, including N degrees, is 2N-1. It is possible to do both continuous and discrete wavelet conversion.

6) Discrete Meyer: It has no explicit statement, it is symmetric, and orthogonal is the analysis. Discrete wavelet transformation can be achieved.

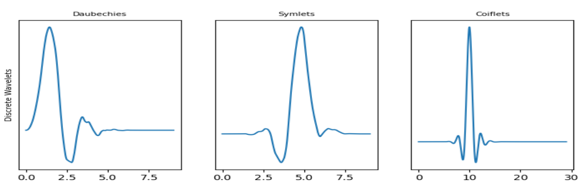

Figure 1 Discrete Wavelets -1

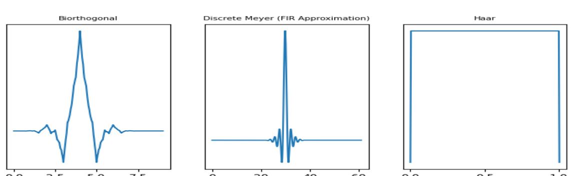

Figure 2 Discrete Wavelets -2

B. Continuous Wavelet Transform

The continuous wavelet transform is obtained by projecting the sign into functions created by the translation and expansion of a main wavelet function . The main wavelet function is a prototype function used to create wavelets. Wavelets are created by the translation and expansion of the main wavelet function. Complex wavelets contains Complex Gaussian Wavelets, Complex Morlet Wavelets, Shannon Wavelets, Frequency B-Spline Wavelets. [3]

$$W_{(a,b)} = \frac{1}{\sqrt{2}} \int \phi\left(\frac{t-b}{a}\right) f(t)dt \tag{5}$$

In the general continuous wavelet transform formula given above, the f(t) function defines the mother wavelet function.

1) Morlet: It is symmetrical, the analysis is not orthogonal or biortogonal. Since there is no scale function, discrete wavelet transform cannot be done.

$$\Psi(t) = \frac{1}{\sqrt{2\pi}} e^{-ifte^{\frac{-t^2}{2}}} \tag{6}$$

In this equation, f represents the frequency of the wavelet.

2) Mexican Hat: It is symmetrical, the analysis is not orthogonal or biortogonal. Since there is no scale function, discrete wavelet transform cannot be done.

$$\Psi(t) = \left(1-t^2\right) e^{\frac{-t^2}{2}} \quad , -\infty < t < +\infty \tag{7}$$

3) Complex Wavelet: Complex wavelets have complicated values. As a consequence, it can simultaneously detect components of phase and amplitude. Therefore, when they are applied to non-stationary signals, these complex-valued wavelets provide better results.

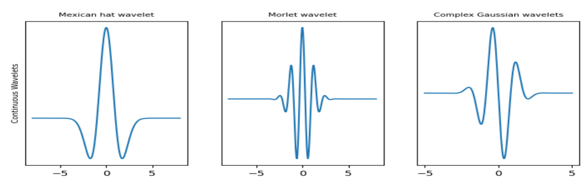

Figure 3 Continuous Wavelets -1

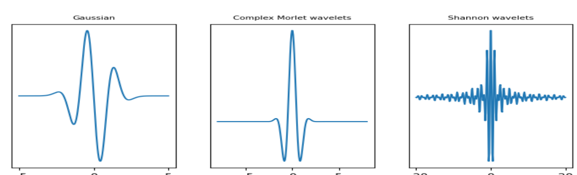

Figure 4 Continuous Wavelets -2

III. MUSIC CLASSIFICATION

A. K Nearest Neighbor Algorithm

KNN (K Nearest Neighbors) is an algorithm used in classification and regression problems. Learning takes place based on the similarity between the data.[4] Similarity calculation is most often done with Euclidean distance.[5]

$$\text{Euclidean Distance:} \qquad d(x,y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \tag{8}$$

B. Support Vector Machine Algorithm

SVM are algorithms frequently used in regression and classification problems. The goal in SVM is to find a hyperplane that can classify input vectors in N dimensional space. SVM algorithms are divided into linear and non-linear SVM algorithms. Linearly separable SVM are used for linearly separable data. Polynomial kernel or Gaussian RBF is used for data that cannot be separated linearly. [6]

C. Random Forest Algorithm

Random Forest Algorithm is classification algorithm and machine learning model that aims to increase the classification value by generating more than one decision tree during the classification process. [7]

D. Naive Bayes Algorithm

Naive bayes method is a supervised learning algorithm based on Bayes theorem. It can be used in unbalanced data sets and can make an effective classification with less data. In this study, Gaussian Naive Bayes algorithm is implemented. [8]

$$\text{Bayes Theorem: } P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{9}$$

$$\text{Gaussian NB: } P(x_i|y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(-\frac{\left(x_i-\mu_y\right)^2}{2\sigma_y^2}\right) \tag{10}$$

IV. Experimental Results

The tables below contain the accuracy, precision and recall informations of the classification algorithms in the dataset processed by discrete wavelet transform. While performing the discrete wavelet transform, 3 levels were selected. The data are written in the tables in the format of 1st level- 2nd level- 3rd level.

Table 1 Discrete Wavelet Functions Average Accuracy Table

Type	Random Forest	KNN	SVC Poly
sym	61.90-65.27-67.25	41.50- 44.88-46.06	42.30- 43.89-50.15
bior	59.41- 59.11-60.30	45.37- 45.18-49.16	46.48- 47.47-54.03
coif	62.18- 64.57-65.78	39.51- 40.40-44.39	43.79- 44.39-48.36
haar	58.01- 58.80-59.70	45.28- 45.07-48.37	44.68- 47.75-53.32
db	58.92- 57.32-60.30	44.28- 45.98-47.55	45.97- 48.17-52.94
rbio	58.41- 58.11-59.82	45.78- 45.07-48.16	45.57- 48.36-53.74
dmey	61.40- 63.87-66.88	43.48- 43.79-46.16	41.48- 42.08-48.66

Table 2 Discrete Wavelet Functions Average Accuracy Table

Type	Gaussian NB	SVC Linear	SVC RBF
sym	38.51 -38.02 -39.30	32.23 -37.81 -44.67	46.66 -50.25-53.75
bior	36.94 -40.70 -39.21	39.00 -48.95 -42.38	51.93 -57.71-54.74
coif	38.50 -39.01 -40.20	33.33 -36.73 -42.29	47.16 -50.45-52.05
haar	36.91 -38.90 -40.79	38.22 -42.39 -48.55	52.54 -52.82-55.13
db	37.91 -39.01 -41.00	38.91 -41.79 -48.16	53.53 -54.33-54.94
rbio	37.41 -40.20 -40.99	38.02 -42.09 -48.76	52.33 -51.85-55.51
dmey	35.83 -35.62 -36.53	34.04 -37.42 -44.98	46.67 -49.66-50.05

Table 3 Discrete Wavelet Functions Average Precision Table

Type	Random Forest	KNN	SVC Poly
sym	64.02- 68.33-69.75	43.11- 46.47-47.93	45.37- 48.15-53.30
bior	61.35- 61.43-62.23	48.88- 46.85-51.10	48.24- 49.76-56.23
coif	64.95- 67.26-68.32	41.90- 42.81-46.42	48.28- 45.43-50.34
haar	59.97- 60.77-60.83	47.00- 47.65-50.83	46.78- 49.59-55.51
db	61.07- 58.73-61.35	46.24- 48.27-49.90	47.21- 50.70-55.82
rbio	60.85- 59.87-62.49	47.35- 47.72-49.85	47.81- 51.65-56.66
dmey	63.79- 65.97-69.25	45.84- 46.67-49.31	43.59- 44.89-51.54

Table 4 Discrete Wavelet Functions Average Precision Table

Type	Gaussian NB	SVC Linear	SVC RBF
sym	36.98- 39.13-39.47	29.17- 35.28-44.09	50.60- 52.91-56.41
bior	35.23- 36.81-40.71	38.01- 43.66-49.53	54.42- 57.46-60.32
coif	36.83- 38.10-38.11	29.76- 35.52-41.65	49.68- 53.37-53.73
haar	34.36- 33.43-41.04	37.88- 40.96-48.37	55.09- 54.57-57.28
db	35.05- 35.59-43.46	37.67- 42.83-49.88	55.53- 55.88-57.64
rbio	34.62- 37.10-43.20	37.93- 42.12-48.92	55.06- 54.97-56.37
dmey	33.41- 34.71-32.18	30.64- 36.27-45.53	48.58- 51.57-52.02

Table 5 Discrete Wavelet Functions Average Recall Table

Type	Random Forest	KNN	SVC Poly
sym	61.90- 65.27- 67.25	41.50- 44.88- 46.06	42.30- 43.89- 50.15
bior	59.41- 59.11- 60.30	45.37- 45.18- 49.16	46.48- 47.47- 54.03
coif	62.18- 64.57- 65.78	39.51- 40.40- 44.39	43.79- 44.39- 48.36
haar	58.01- 58.80- 59.70	45.28- 45.07- 48.37	44.68- 47.75- 53.32
db	58.92- 57.32- 60.30	44.28- 45.98- 47.55	45.97- 48.17- 52.94
rbio	58.41- 58.11- 59.82	45.78- 45.07- 48.16	45.57- 48.36- 53.74
dmey	61.40- 63.87- 66.88	43.48- 43.79- 46.16	41.48- 42.08- 48.66

Table 6 Discrete Wavelet Functions Average Recall Table

Type	Gaussian NB	SVC Linear	SVC RBF
sym	38.51 -38.02- 39.30	32.23- 37.81- 44.67	46.66- 50.25- 53.75
bior	36.94 -40.70- 39.21	39.00- 48.95- 42.38	51.93- 57.71- 54.74
coif	38.50 -39.01- 40.20	33.33- 36.73- 42.29	47.16- 50.45- 52.05
haar	36.91 -38.90- 40.79	38.22- 42.39- 48.55	52.54- 52.82- 55.13
db	37.91 -39.01- 41.00	38.91- 41.79- 48.16	53.53- 54.33- 54.94
rbio	37.41 -40.20- 40.99	38.02- 42.09- 48.76	52.33- 51.85- 55.51
dmey	35.83 -35.62- 36.53	34.04- 37.42- 44.98	46.67- 49.66- 50.05

As can be seen from the tables, it is the Random Forest classification algorithm with the highest accuracy. This is because of the algorithm creates many decision trees and obtains an efficient result by following the most appropriate paths. Following the Random Forest algorithm is SVC RBF (Radial Basis Function) and SVC Poly algorithms. SVC RBF and SVC Poly, which are nonlinear SVC algorithms, perform the classification process with kernel functions in data sets where a linear hyperplane cannot be drawn. Since the Linear SVC algorithm works with high accuracy on linear data sets, it offers a lower accuracy compared to other DVM algorithms.

Table 7 Continuous Wavelet Functions Average Accuracy Table

Type	Random Forest	KNN	SVC Poly
cgau	51.74	38.11	42.50
cmor	51.65	35.22	37.91
gaus	53.02	38.30	42.10
fbsp	49.95	36.41	38.01
morl	54.91	38.70	38.21
mexh	52.35	37.11	42.19
shan	54.62	33.73	39.21

Table 8 Continuous Wavelet Functions Average Accuracy Table

Type	GaussianNB	SVC Linear	SVC RBF
cgau	39.21	39.51	48.37
cmor	35.43	39.11	44.97
gaus	39.71	39.70	48.56
fbsp	35.23	38.52	45.16
morl	35.41	37.51	44.58
mexh	37.02	38.80	44.77
shan	37.81	39.21	46.29

Table 9 Continuous Wavelet Functions Average Precision Table

Type	Random Forest	KNN	SVC Poly
cgau	52.95	40.59	44.91
cmor	53.99	37.77	42.45
gaus	55.14	42.48	44.77
fbsp	50.55	39.07	40.40
morl	57.94	41.93	40.48
mexh	53.31	39.26	44.16
shan	56.47	35.18	42.59

Table 10 Continuous Wavelet Functions Average Precision Table

Type	GaussianNB	SVC Linear	SVC RBF
cgau	38.22	37.85	50.45
cmor	34.34	36.94	46.73
gaus	42.76	39.50	50.17
fbsp	35.95	35.41	46.76
morl	31.96	35.62	46.91
mexh	39.49	38.24	44.58
shan	38.53	36.97	48.63

Table 11 CONTINUOUS WAVELET FUNCTIONS AVERAGE RECALL TABLE

Type	Random Forest	KNN	SVC Poly
cgau	51.74	38.11	42.50
cmor	51.65	35.22	37.91
gaus	53.02	38.30	42.10
fbsp	49.95	36.41	38.01
morl	54.91	38.70	38.21
mexh	52.35	37.11	42.19
shan	54.62	33.73	39.21

Table 12 CONTINUOUS WAVELET FUNCTIONS AVERAGE RECALL TABLE

Type	GaussianNB	SVC Linear	SVC RBF
cgau	39.21	39.51	48.37
cmor	35.43	39.11	44.97
gaus	39.71	39.70	48.56
fbsp	35.23	38.52	45.16
morl	35.41	37.51	44.58
mexh	37.02	38.80	44.77
shan	37.81	39.21	46.29

Once again, the highest performance is obtained with the Random Forest algorithm. Next comes SVC RBF and SVC Poly algorithms. Results produced with Morl family and Shannon family have the highest values of accuracy, precision and recall. As seen in the tables, the wavelet families that give the most successful percentages differ according to the different classification algorithms selected. Additionally, looking at the recall and precision tables, it is not much difference from the accuracy table.

V. CONCLUSION

Within the scope of the project, a data analysis tool has been developed that processes the signals with the wavelet transformation method, passes them through statistical functions and enables classification with machine learning algorithms.

Signals were analyzed by discrete and continuous wavelet transform. It has been determined that discrete wavelet transform spends less time than continuous wavelet transform and produces more accurate results.

The highest accuracy (%67.25) in music classification was obtained in the sym wavelet family with the Random Forest algorithm, and at the 3rd level with the discrete wavelet transform.

For continuous wavelet transformation, the most successful result (%54.91) was obtained by using the Random Forest algorithm with the morl wavelet family.

It has been determined that the more the level parameter is increased in the discrete wavelet transform, the more successful results are obtained. For example, the classification of the signal processed with the sym wavelet family and applied Random Forest algorithm has an accuracy rate of %61.90 at level 1, %65.27 at level 2 and %67.25 at level 3.

REFERENCES

[1] R. Polikar, "The engineer's ultimate guide to wavelet analysis, the wavelet tutorial," 2006. [Online]. Available: https://cseweb.ucsd.edu/ baden/Doc/wavelets/polikar$_{w}avelets.pdf$

[2] "Pywavelets - wavelet transforms in python." [Online]. Available: https://pywavelets.readthedocs.io/en/latest/index.html

[3] ThePyWaveletsDevelopers. (2016) Continuous wavelet transform (cwt). [Online]. Available: https://pywavelets.readthedocs.io/en/latest/ref/cwt.html

[4] H. Vinchhi. Machine learning : Introduction to k nearest neighbor (knn). [Online]. Available: http://assignmentsolutionguru.com/article/intro-to-knn-in-python

[5] S. A. Gokte. Most popular distance metrics used in knn. [Online]. Available: https://www.kdnuggets.com/2020/11/most-popular-distance-metrics-knn.html

[6] N. S. Chauhan. A friendly introduction to support vector machines. [Online]. Available: https://www.kdnuggets.com/2019/09/friendly-introduction-support-vector-machines.html

[7] W. Koehrsen. Random forest simple explanation. [Online]. Available: https://williamkoehrsen.medium.com/random-forest-simple-explanation-377895a60d2d

[8] scikit learn. Naive bayes. [Online]. Available: https://scikit-learn.org/stable/modules/naive$_{b}ayes.html$