2017 IEEE INTERNATIONAL WORKSHOP ON MACHINE LEARNING FOR SIGNAL PROCESSING, SEPT. 25–28, 2017, TOKYO, JAPAN

OBJECT CLASSIFICATION WITH CONVOLUTION NEURAL NETWORK BASED ON THE TIME-FREQUENCY REPRESENTATION OF THEIR ECHO

Mariia Dmitrieva, Matias Valdenegro-Toro, Keith Brown, Gary Heald, David Lane

Heriot-Watt University, Edinburgh, UK

ABSTRACT

This paper presents classification of spherical objects with different physical properties. The classification is based on the energy distribution in wideband pulses that have been scattered from objects. The echo is represented in Time-Frequency Domain (TFD), using Short Time Fourier Transform (STFT) with different window lengths, and is fed into a Convolution Neural Network (CNN) for classification. The results for different window lengths are analysed to study the influence of time and frequency resolution in classification. The CNN performs the best results with accuracy of $(98.44 \pm 0.8)\%$ over 5 object classes trained on grayscale TFD images with 0.1 ms window length of STFT. The CNN is compared with a Multilayer Perceptron classifier, Support Vector Machine, and Gradient Boosting.

***Index Terms*—** Wideband pulses, time-frequency representation, convolution neural networks, object classification

1. INTRODUCTION

Most sonar data target classification approaches make use of 2D and 3D images. This classification is based on the object's shape. For applications such as bio-acoustical analysis, underwater archaeology, oil and gas pipe maintenance it is required to know physical properties of objects. In most cases object's shape doesn't provide sufficient information to discriminate these different properties. The study of the scattering using a wide frequency range provides the opportunity to gain more information.

The wideband pulses are used for different tasks including object identification and classification. Gaunaurd et al. [1] insonify a sphere with recorded wideband dolphin pulses and define specific features of the echoes connected to the physical properties of the sphere. Pailhas et al. [2] classify seven different objects based on the location of the notches in Frequency Representation of their wideband echoes. Qiao et al. [3] apply Singular Value Decomposition for echoes in Time-Frequency Domain (TFD) and use Support Vector Machine to classify three copper cylindrical shells with different thickness. Neural Networks were applied for prediction of an acoustic Form Function of an infinite length stainless steel tube by Dariochy et al. [4].

In contrast to the previous approaches [1–4], in this work we apply Convolution Neural Network (CNN) to classify objects with different physical properties based on the Time-Frequency Representation (TFR) of their scattering pulses. All the objects have a spherical shape and differ in the material of their shell, filler material, shell thickness and diameter. The spherical shape of the objects is beneficial when focussing on their physical properties. Firstly, scattering from a sphere doesn't depend on the view angle. Secondly, when all the objects have similar spherical shape one can consider difference in the scattering influenced mostly by physical properties of the objects.

The targets are submerged in water and insonified by a wideband pulse. The pulses are single linear chirps with constant frequency range from 30 kHz to 160 kHz. The duration of the pulses is varied from $0.1ms$ to $2.0ms$. The scattered pulse in recorded in the Time Domain (TD). The Fourier Transform allows us to decompose a signal into frequency components and present the scattering in Frequency Domain (FD). Signals also can be represented in Time-Frequency Domain (TFD) where the energy of the signal is distributed in Time and Frequency. In this work, we use Short Time Fourier Transform (STFT) to present the echo in TFD for classification. Images of the echo in TFD are fed to a Convolution Neural Network.

Different window length of STFT are applied to investigate the influence of the time and frequency resolution into the performance of the classifier. Results from the CNN classifier are compared with results of Multilayer Perceptron (MLP), Support Vector Machine (SVM) and Gradient Boosting (GB). In this work we show that spherical objects can be classified with high accuracy based on the TFR images of their echo with a CNN classifier.

Acquisition of the data is presented in Section 2. Section 3 describes preprocessing of the recordings and Time-Frequency Representation of the echoes. In Section 4 CNN and MLP classifiers are presented. Results are described in Section 5 with conclusion and future work in Section 6.

Thanks to the FP7-PEOPLE-2013-ITN project ROBOCADEMY (Ref 608096) for funding.

978-1-5090-6341-3/17/$31.00 ©2017 IEEE

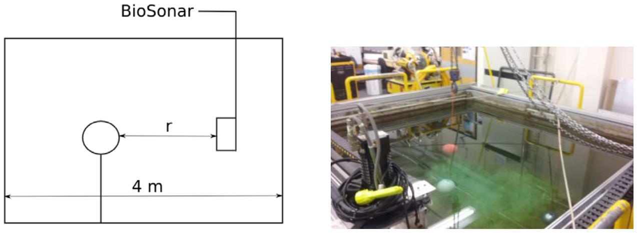


Fig. 1. Experimental set-up.

2. SIGNAL RECORDING

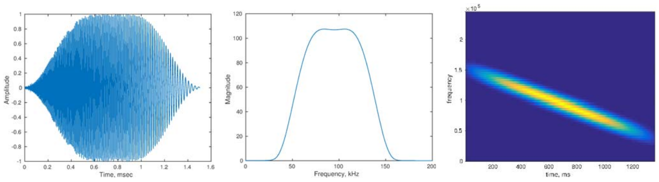
2.1. Data Acquisition

The echos are recorded in a $3m \times 4m \times 2m$ water tank using the Hydrason wideband sonar (Biosonar), Figure 1. The objects are fixed in the water using weight and placed in front of the BioSonar. The object is about $1.5m - 3m$ away from the sonar. The wideband sonar works in the frequency range $30kHz$ to $160kHz$ and allows transmitting pulses of different shape and duration.

2.2. Pulse Configuration

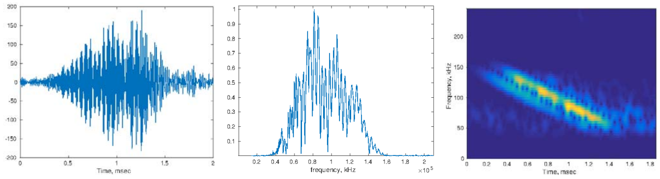
In this work we use linear chip pulses $30-160kHz$, Figure 2. The pulse duration is varied from $0.1ms$ to $2.0ms$ in $0.1ms$ increments.

The echo changes from an identical replica of the initial pulse due to the scattering from the object. The pulse shape and frequency composition of the echo transform due to the interaction with the object. These changes expose properties of the object, including size, shape, material and structure. Figure 3 illustrates the changes. It presents the pulse reflected from a spherical aluminium object filled with water. The scattered echo has a visible differences compared to the original pulse.



(a) TD (c) FD (c) TFD

Fig. 2. Linear chirp in range 30-160 kHz, duration 1.5 ms.



(a) TD (c) FD (c) TFD

Fig. 3. Scattering of a linear chirp in range 30-160 kHz, duration 1.5 ms.

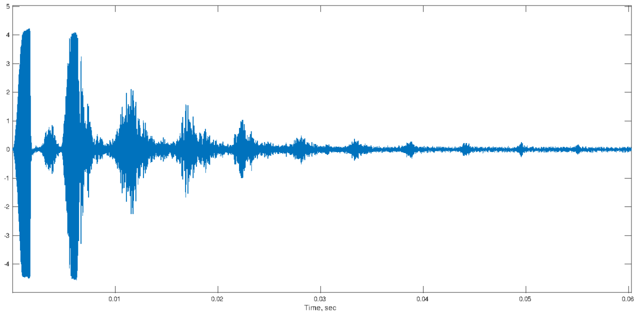


Fig. 4. Recording of a response

3. ECHO REPRESENTATION

The recorded signals are processed in two steps. First the echo is selected from the recording and then it is represented in Time-Frequency Domain. The image of the echo in the TFD is used for the classification.

3.1. Echo Selection

Recordings made in a water tank contains reflections from the object, walls, bottom of the tank and other surfaces, Figure 4.

The echo segment from the object contains the only useful information for the classifier. The segment is selected from the recording based on the matched filtering of the initial and returned pulses. The surfaces the pulse reflected from are presented as peaks on the output of the filter, Figure 5.

The peak related to the object of interest is located in the known range between $1.5m$ and $3m$, corresponding to time of $0.002ms$ and $0.004ms$. The first significant peak in the range is chosen as the target's surface. The starting point of

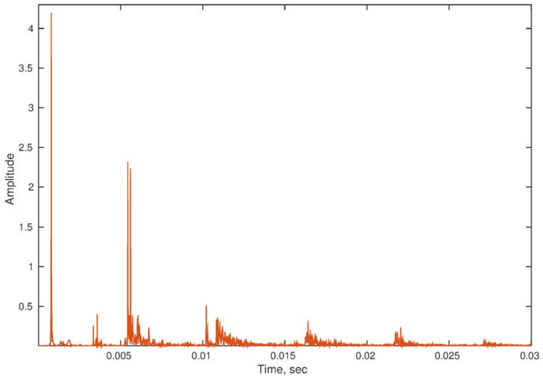


Fig. 5. Matched filter's response

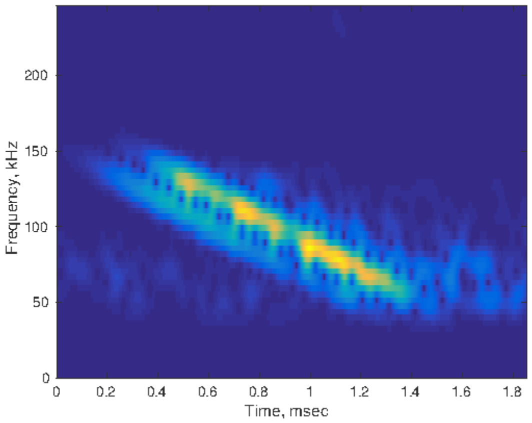


Fig. 6. Echo from an aluminium sphere filled with water presented by a magnitude of the STFT.

the scattering, t_{sc}, is calculated based on the peak position, t_{peak}, Equation 1.

$$t_{sc} = t_{peak} - \Delta t_{pulse}/2, \tag{1}$$

where Δt_{pulse} is the initial pulse duration. The duration of the scattering sample is fixed for all experiments and equals $2ms$. The value is chosen based on the duration of the initial pulses and geometry of the water tank.

The segmented echo is then transformed into Time-Frequency Domain.

3.2. Time Frequency Representation

The Time-Frequency Representation (TFR) of a signal provides distribution of its energy over time and frequency simultaneously [5]. It describes how a signal's frequency composition changes with time. There are many different techniques which transform a signal from Time into the Time Frequency Domain [6].

In this work we use magnitude of the Short Time Fourier Transform (STFT). The signal in Time Domain is segmented into a sequence of short sections using a window function $w(n)$. Each segment is transformed into the frequency domain using a Fourier Transform. The magnitude of the signal can be displayed as 2D plot with frequency and time axis, Figure 6. The STFT images are presented in colour for visualisation purposes, while for classification single channel grayscale images were used. The value of each point on the graph represents an energy of a particular time interval and frequency.

STFT of a discrete signal $x(n)$ can be presented by Equation 2, [7].

$$X_m(\omega) = \sum_{n=-\infty}^{\infty} x(n)w(n-mR)e^{-jwn}, \tag{2}$$

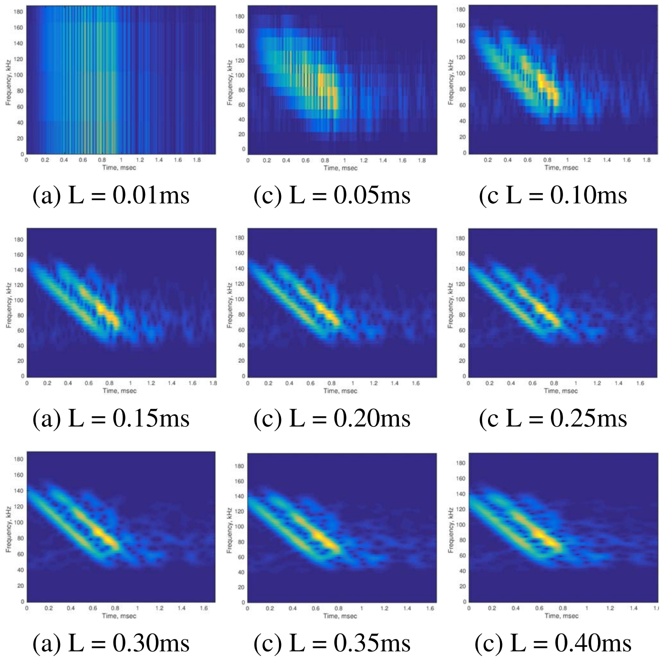

(a) L = 0.01ms
(c) L = 0.05ms
(c L = 0.10ms
(a) L = 0.15ms
(c) L = 0.20ms
(c L = 0.25ms
(a) L = 0.30ms
(c) L = 0.35ms
(c) L = 0.40ms

Fig. 7. Magnitude of the STFT with different window length L.

where R is a step size, in samples, between successive DTFTs.

The STFT doesn't contain a cross-product, which is common for the Wigner-Ville Distribution, but has its peculiarity. The approach is a trade-off between time and frequency resolution. The resolution in time is evaluated by the length of the window, L. A segment with a length L is presented by a single vector on frequency axis. Decreasing the window length L leads to improvement in time resolution, Δt.

The frequency resolution depends on the sampling frequency f_s and number of samples in the window N, Equation 3.

$$\Delta f = \frac{f_s}{N} \tag{3}$$

Increasing the window length improves the frequency resolution. In this way the window length is a compromise between resolutions of time and frequency axis.

The length of the scattering segment equals $2ms$ for all the experimental data with sampling frequency 1 MHz. Figure 7 illustrates influence of the STFT window length on the time and frequency resolution. The signal is an echo from an aluminium object. The initial pulse is a $1.5ms$ linear chirp with a frequency range from 30 to $160kHz$.

The window length, L, is varied from from $0.01ms$ to $0.4ms$, while the step size, R, is kept constant. It brings variation in frequency resolution from $100kHz$ to $2.5kHz$. Based on the visual observation of the TFR images we choose $0.15ms$ as major window length for the classifier. It presents a good compromise for the time and frequency resolution. The CNN is built for the example of 0.15 ms window length and then trained and tested for all the other values as well.

4. OBJECT CLASSIFICATION

4.1. Classes

The echo is classified into five classes with four objects and one empty scene, Table 1. The empty scene doesn't contain any objects within the expected range from $1.5m$ to $3m$.

Table 1. Description of the objects we are interested in identifying.

	object	diameter, m	shell material	filler
1	sphere	0.15	aluminium	water
2	sphere	0.274	plastic	air
3	sphere	0.208	plastic	water
4	sphere	0.208	plastic	air
5	no objects	-	-	-

The classified objects have similar spherical shapes and differ by radius, thickness of the shell, filler and shell materials. These parameters of a spherical target influence the scattering and create specific changes in the initial pulse. The spherical shape of the objects eliminates influence of the view point into the scattering.

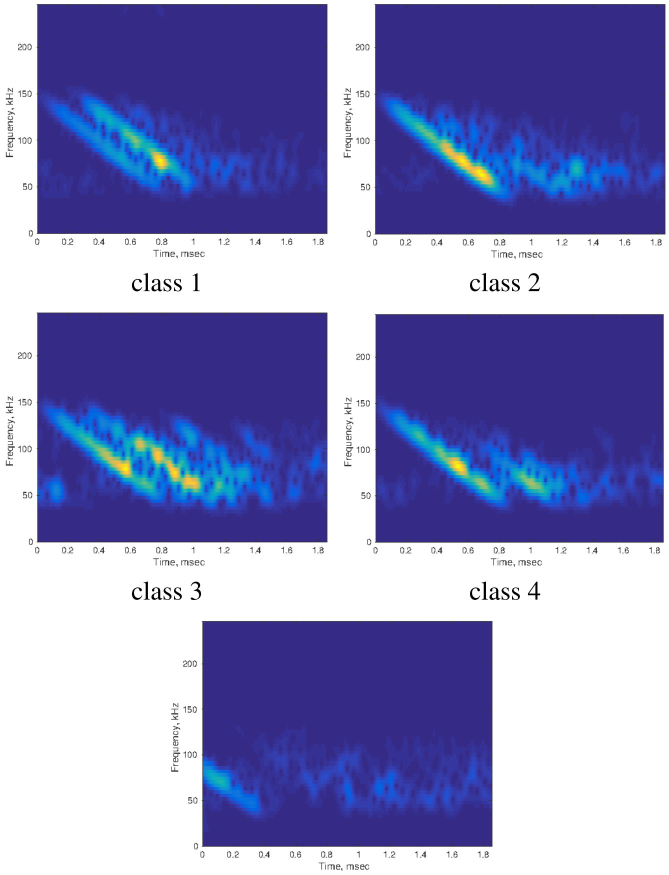
Each object is insonified with linear chirp pulses. Duration of the pulses changes from $0.1ms$ to $2ms$. Figure 8 illustrates scattering of $1ms$ pulse in TFD for each class. STFT is implemented with window length $0.15ms$.

4.2. Classification Models

The use of different neural network classifiers, namely a Convolutional Neural Network, and a Multilayer Perceptron (MLP) is evaluated and compared with other classifiers. Recent results that show that CNNs are one of the best classifiers for image data, this is the motivation for this work.

We use the following notation. Conv2D(n, s) a 2D Convolutional layer with n square filters of size s, Max-Pool(s) a Max-Pooling layer with subsampling size s, FC(n) a fully connected layer with n output neurons is used. The following networks were designed:

- A CNN with configuration Conv2D(8, 5×5)-MaxPool(2×2)-Conv2D(8, 5 × 5)-MaxPool(2 × 2)-FC(32)-FC(5). This network takes a 50x75 input image. The network has 36973 parameters in total for a three-channel RGB colour image input, while the network has 36573 parameters for one-channel grayscale image. The network architecture is shown in Fig. 9.
- An MLP with configuration FC(64)-FC(64)-FC(5). The input to this network is a flattened image vector of size $c \times 50 \times 75$, where c is the number of channels in the input image. This network has 724549 parameters for a three-channel RGB colour image. The model architecture is shown in Fig. 10



class 1

class 2

class 3

class 4

class 5

Fig. 8. TFR of the scattering for each class.

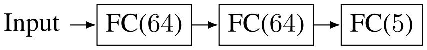


Fig. 10. Multilayer Perceptron Architecture

To prevent overfitting, FC layers are followed by Dropout layers [8] with $p = 0.5$, except when they are output layers. ReLU activations (Eq. 4) are used through the network, except for output layers that use the Softmax activation (Eq. 5) instead.

$$f(x) = \max(0, x) \tag{4}$$

$$f(\mathbf{x}) = \left[\frac{e^{x_i}}{\sum_j e^{x_j}} \right]_i \tag{5}$$

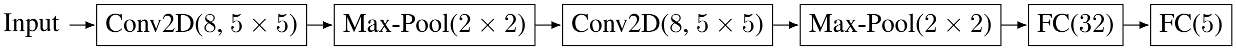


Fig. 9. Convolutional Neural Network Model Architecture. All layers use ReLU activation, except the last layer that uses a softmax function.

The networks are trained with Stochastic Gradient Descent (SGD) for 15 epochs with a learning rate $\alpha = 0.01$ and a batch size $B = 128$. The loss function that is minimized during training is the Categorical Cross-Entropy:

$$L(y, \hat{y}) = -\sum_{i=0}^{N} \sum_{c=0}^{C} y_i^c \log \hat{y}_i^c, \tag{6}$$

where N is the number of elements in the dataset, and C is the number of classes ($C = 5$ in this work).

5. RESULTS

5.1. Experimental Setup

The experiments were done with a dataset consisting of 5 classes, with 15000 examples per class. In order to provide a fair evaluation, a 5-Fold cross-validation was performed, and at each fold, 5 instances of the same neural network were trained, with different random initializations of their weights. This includes the effect of both data and random initialization on classification accuracy.

25 network instances for each window length parameter were trained. This parameter was selected in the range $L \in [0.01, 0.05, 0.10, 0.20, 0.25, 0.30, 0.35, 0.40]$. For each value of L the mean and standard deviation of accuracy over the 25 trained networks (5 folds times 5 network instances) are reported. The results are summarized in Table 2.

Two comparisons are performed. The first is the variation of the window length parameter L. The second is the use of a CNN, MLP, SVM, and Gradient Boosting classifiers. We trained classifiers using the same cross-validation methodology and report mean and standard deviation of test accuracy. We used a linear SVM with regularization coefficient $C = 10$, and a Gradient Boosting classifier with 100 stages, learning rate $\alpha = 0.1$, and maximum tree depth of seven. Both classifier configurations were obtained by grid search on a validation part of the dataset.

5.2. Window length

CNN classifier results for TFD images are presented in Figure 11. The highest accuracy of (98.44 ± 0.84) % is achieved for window length $0.1ms$. Our results show that this window length provides a clear visual representation of the scattering with a fine compromise of time and frequency resolutions.

These results show that a high accuracy classification can be achieved either for a fine trade off the time and frequency resolutions, either for a high resolution in a single time or frequency domain.

Standard deviation decreases towards shortest and longest window lengths. It can be interpreted as a better and more stable model fit to the data for these window length.

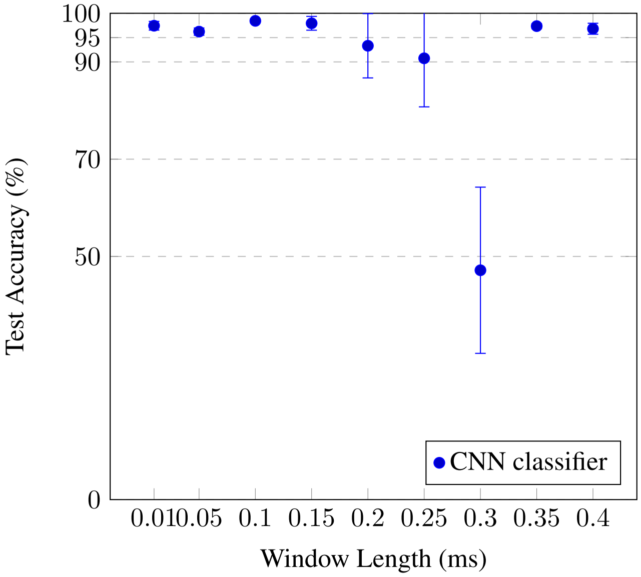


Fig. 11. Accuracy of the CNN classifiers for different window length.

5.3. Classifier Comparison

Figure 12 shows our results using CNN, MLPs and other classifiers. Table 2 shows a numerical view of our results.

Our results show that CNNs perform considerably better than MLPs for all variations of the window length L. MLPs perform poorly, with the highest accuracy close to 60%. All models obtain considerably lower performance at $L = 0.30ms$, but classification performance increases after that point. It can be assumed that the drop in performance is caused by a deterioration of resolution in time and frequency.

But other classifiers get similar performance, but not better than a CNN. An SVM outperforms our CNN for $L = 0.20$, $L = 0.25$ and $L = 0.30ms$ window length. This indicates that a SVM is able to generalize where a CNN cannot, specially for the case of $L = 0.30ms$. We believe that this is a mild indication of overfitting, as a SVM is not as restricted as a CNN as how the input image is interpreted and features extracted. A single pixel could be used by an SVM to produce

Table 2. Classification Performance of CNN, MLPs and other classifiers for a Range of Window Lengths L.

L, ms	0.01	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40
Δf. kHz	100	20	10	6.7	5	4	3.3	2.8	2.5
Mean Accuracy ± Standard Deviation, %									
CNN	**97.4 ± 1.8**	**96.2 ± 1.5**	**98.4 ± 0.8**	**97.9 ± 2.8**	93.3 ± 13.2	**90.7 ± 19.9**	47.2 ± 34.2	**97.4 ± 1.0**	**96.8 ± 2.3**
MLP	60.3 ± 18.1	55.8 ± 15.4	51.3 ± 17.8	57.2 ± 17.7	46.4 ± 17.1	46.9 ± 16.0	51.4 ± 21.5	55.6 ± 15.5	59.1 ± 14.4
SVM	96.9 ± 0.04	95.8 ± 0.14	98.4 ± 0.1	**98.5 ± 0.02**	**98.9 ± 0.03**	**98.0 ± 0.12**	**98.8 ± 0.14**	96.94 ± 0.2	96.0 ± 0.12
GB	94.9 ± 0.08	94.0 ± 0.24	97.3 ± 0.2	97.3 ± 0.23	97.2 ± 0.14	96.7 ± 0.18	97.4 ± 0.24	94.7 ± 0.12	94.7 ± 0.14

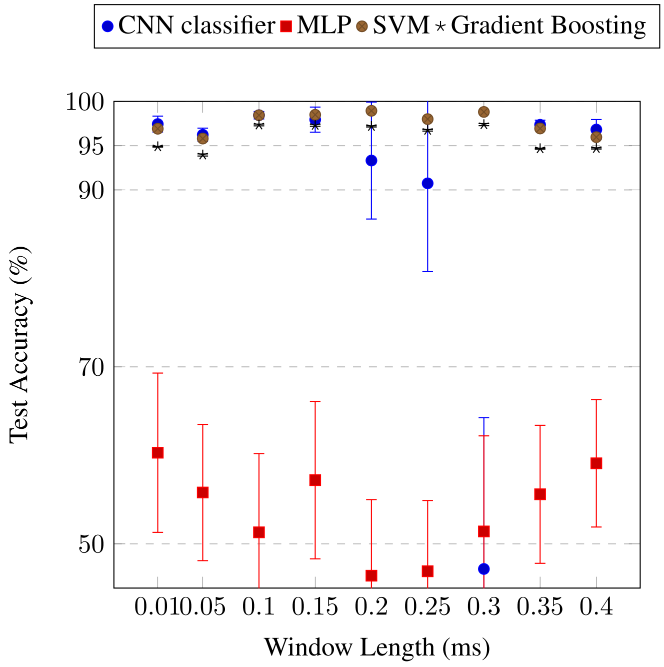


Fig. 12. Accuracy of the CNN and MLP classifiers.

a class decision.

Gradient boosting also performs well but does not outperform both a CNN or SVM.

6. CONCLUSIONS AND FUTURE WORK

This work presents a CNN classifier of spherical objects with different physical properties based on TFR of their scattering. The highest performance of CNN classifier is achieved with accuracy $(98.44 \pm 0.8)\%$ for grayscale TFR images with window length $L = 0.10ms$. It is shown that a TFR image is a reasonable descriptor, which allows to achieve high classification accuracy. The results illustrate that CNN classifier is outperform MPL and GB for all range of the window lengths. SVM outperforms CNN in a number of window lengths, which can be a mild indication of overfitting.

It is observed that by changing the STFT window length one can achieve high performance classification in two cases: finding compromise of time and frequency resolution based on the visual representation or picking a high resolution in a single time or frequency domain

The effect of the STFT window length on the classification results requires further study. The influence of the time and frequency resolution can depend on the initial pulse and object properties. For future work, we would like to generalize our results for bigger datasets and larger number of objects, with varying physical properties.

7. REFERENCES

[1] Guillermo C. Gaunaurd, Donald Brill, Hanson Huangb, Patrick W. B. Moore, and Hans C. Strifors, "Signal processing of the echo signatures returned by submerged shells insonified by dolphin clicks: active classificatio," *The Journal of the Acoustical Society of America*, vol. 103(3), pp. 1547–1557, 1998.

[2] Y. Pailhas, C. Capus, K. Brown, and P.W. Moor, "Analysis and classification of broadband echoes using bio-inspired dolphin pulses," *The Journal of the Acoustical Society of America*, vol. 127, pp. 3809–3820, 2010.

[3] Gang Qiao, Xin Qing, Donghu Nie, Yi Zhang, and Jiangsheng Tanh, "Underwater cylindrical shell in different thickness recognition using biomimetic dolphin clicks," *Proceedings of Ocean'16*, 2016.

[4] A. Dariouchy, E. Aassif, G. Maze, D. Decultot, and A. Moudden and, "Prediction of the acoustic form function by neural network techniques for immersed tubes," *The Journal of the Acoustical Society of America*, vol. 124, 2008.

[5] Leon Cohen, "Time-frequency distirbution - a review," *Proceesings of the IEEE*, 1989.

[6] Boualem Boashash, *Time-Frequency Signal Analysis and Processing. A comprehensive Reference.*, Elsevier, 2016.

[7] J.B. Allen and L.R. Rabiner, "A unified approach to short-time fourier analysis and synthesis," *Proceesings of the IEEE*, vol. 65, pp. 1558–1564, 1977.

[8] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.