Uluslararası İleri Doğa Bilimleri ve Mühendislik Araştırmaları Dergisi
Sayı 8, S. 272-278, 4, 2024

Araştırma Makalesi

IJANSER
https://as-proceeding.com/index.php/ijanser
ISSN: 2980-0811

International Journal of Advanced Natural Sciences and Engineering Researches
Volume 8, pp. 272-278, 4, 2024

Research Article

Comparing the Performance of Deep Learning Architectures for Sentiment Analysis

Simge YILDIRIM*, Yunus SANTUR [2]

[1]*Software engineering / institute of science, Firat University, Turkey*
[2]*Department of Artificial Intelligence and Data Engineering/ institute of science, Firat University, Turkey*

(yildirimsimge8@gmail.com) Email of the corresponding author

(Received:15 May 2024, Accepted: 25 May 2024)

(3rd International Conference on Engineering, Natural and Social Sciences ICENSOS 2024, May 16-17, 2024)

ATIF/REFERENCE: Yıldırım, S. & Santur, Y. (2024). Comparing the Performance of Deep Learning Architectures for Sentiment Analysis. *International Journal of Advanced Natural Sciences and Engineering Researches*, 8(4), 272-278.

Abstract – With the advancement of technology, people frequently express their feelings and thoughts in environments such as social media. Natural language processing applications are also very much on the agenda. Thanks to sentiment analysis, inferences can be made by analyzing them. In this study, different methods of emotion classification with deep learning were investigated and applied. IMDb dataset created from movie reviews was used as a dataset. In sentiment classification, four different architectures were applied to the same dataset and compared. As a result of this comparison, it was observed that the 1D CNN model gave the best results. It was concluded that this architecture is efficient and fast for such studies.

Keywords – Deep learning, LSTM (Long Short-Term Memory), RNN (Recurrent Neural Network), 1D CNN (Convolutional Neural Network), Emotion Classification

I. INTRODUCTION

This Deep learning is an important subfield of artificial intelligence and machine learning, which is very popular today. These methods are generally developed with a focus on artificial neural networks.

Deep learning involves multilayer computational models to represent the available data at different levels of abstraction [1]. In deep learning techniques, there are successive layers for the representation of data. In these methods, also known as deep networks, high-level abstractions are effectively made from raw data and feature sets are automatically generated. In this way, attributes that are usually determined by humans are automatically identified and made available [2]. It is known that the training phase of some deep learning algorithms takes quite a long time and semi-supervised learning methods have been proposed in various studies to shorten the training time of deep neural networks [3]. In some researches, various approaches have been proposed to bring the capabilities of deep learning to non-deep but widely used machine learning methods such as support vector machines [4]. Deep learning techniques have achieved highly effective results in analyzing various data formats such as video, audio and text [5]. Generally, one method may be effective in areas such as text data processing, while a different approach may give better results in analyzing video and audio data [6]. In some researches, the proposed deep learning techniques can be successfully applied to multi-modality (different modes such as text, image, audio) [7]. Language modeling and natural language processing, speech and audio processing, information retrieval, object recognition and computer vision, multimodal and multitask learning are among the application areas of

deep learning methods that can be discussed in detail [8]. Natural language processing encompasses various automatic operations on human written text, including text analysis, classification, information extraction and sentiment analysis. Its application areas are diversifying and expanding day by day.

Although there are different areas of natural language processing, sentiment analysis is the most widely used. Sentiment analysis has an important place in analyzing for users. The popularity of this field is increasing day by day. In this study, sentiment analysis was performed with users' movie reviews. Haque et al. (2019) compared ESA and Long Short-Term Memory (LSTM) networks on the IMDb dataset and stated that the ESA network gave better results [9]. Rao et al. (2018) proposed a new LSTM model on IMDb and Yelp datasets [10]. Islam et al. (2018) applied Random Forest classifier to IMDb and Amazon datasets [11]. Narayanan et al. (2013) worked with Naive Bayes algorithm on IMDb dataset [12]. Huang et al. (2018) studied LSTM and BiLSTM and concluded that BiLSTM is successful [13].

Pang, et al. (2002) tested different machine learning techniques on the IMDb dataset [14]. Matsumoto et al. (2005) applied machine learning classifiers on IMDb and Polarity datasets [15]. Arzu and Aydogan (2023) implemented and compared Transformers-based architectures [16].

II. Materials and Method

In this section of the study, information about the dataset is given. IMDb dataset was used in the study. This data, which was extracted from Kaggle [17], was first subjected to basic preprocessing and then sentiment analysis was performed with deep learning and the results were compared. The flow diagram of the applied method is as shown in Figure 1.

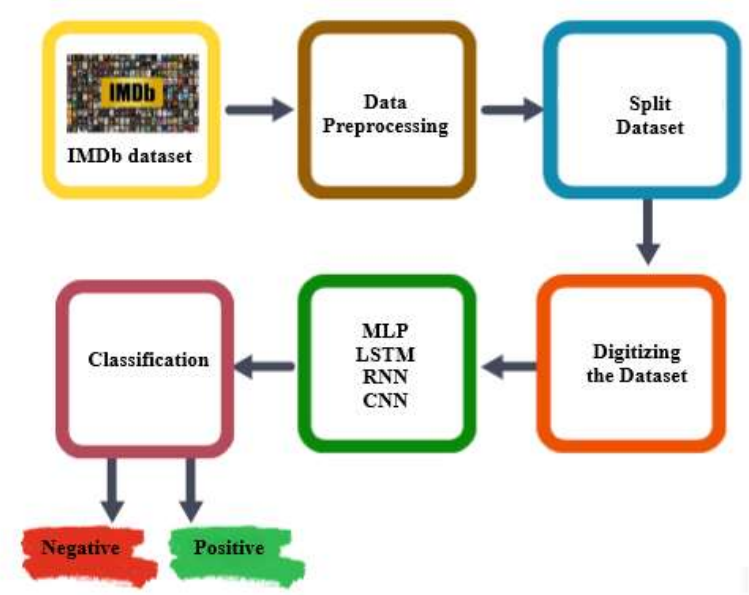

Fig. 1 Emotion Classification Flow Diagram

A. *Dataset*

The IMDb dataset, which is one of the most widely used datasets, was used in this study by conducting the necessary literature review. Although machine learning algorithms are generally used in the literature, deep learning algorithms are used in this study.

IMDb dataset has 50,000 movie reviews. It provides 25,000 movie reviews for training and 25,000 movie reviews for testing. In the dataset, positive tags are used for positive reviews and negative tags are used for negative reviews. In this study, we predict positive and negative reviews using deep learning algorithms. Two examples of positive and negative reviews from the dataset used in the study are shown in Table 1 with their labels to give the reader an idea. Since the comments are in English, their Turkish equivalents are given in parentheses in the table for better comprehension.

Table 1. Sample comments

Dataset	Review	Sentiment
IMDb	Off all the films I have seen, this one. The Rage has got t be one of the worst yet. The direction …	0
	I cannot believe I enjoyed this as much as I did The Anthology stories were better than par, but th…	1

In this study, four different models were created for sentiment analysis prediction using different deep learning models: CNN, RNN, LSTM and MLP. The models were prepared with Python programming language in Jupyter Notebook environment. The performance of the models was evaluated by making predictions on the models. Figure-2 shows the distribution of positive and negative comments in the dataset. As can be seen from Figure 2, the dataset consists of an equal number of positive and negative reviews.

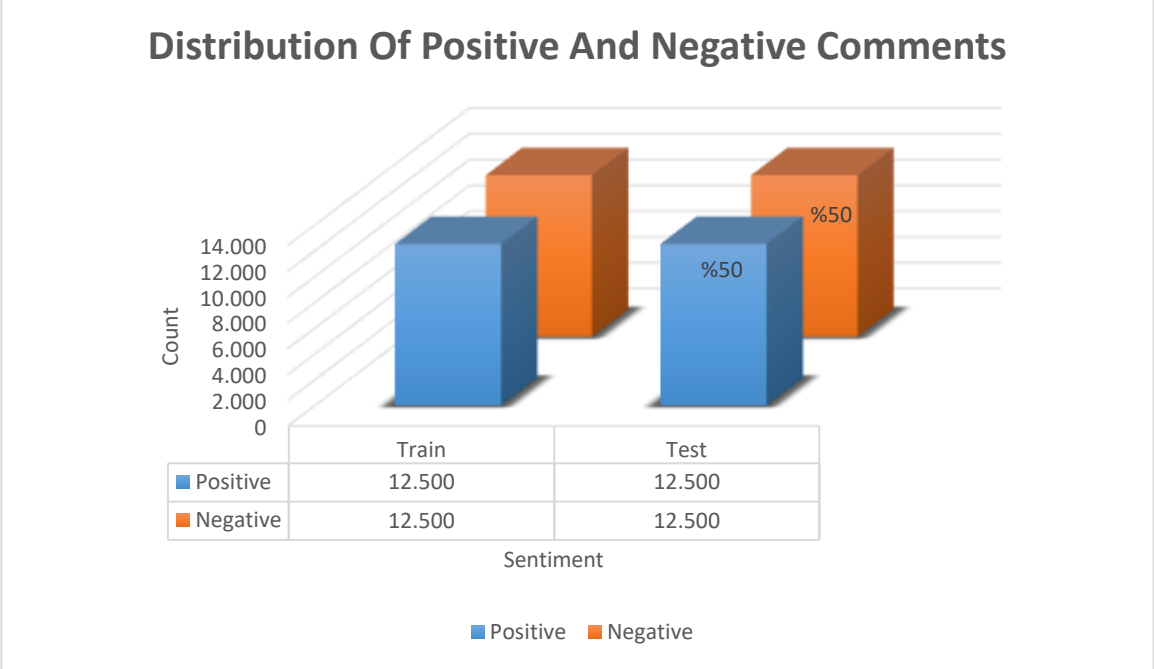

Fig. 2 Distribution of positive and negative comments

B. *Data Pre-Processing*

It is known that data collected for scientific research is usually included in data sets without any pre-processing, i.e. without structuring [18]. However, there is a correlation between the success in the data set and the accuracy of the results obtained from the algorithms used in the data set. In data sets that are not subjected to any pre-processing, both the analysis and the result are badly affected. For this reason, the dataset was preprocessed first. Since HTML tags, special characters and symbols are not helpful in identifying an emotion, we removed them and obtained clean data. Afterwards, all comments were converted to lower case to make it easier to analyze the comments. Although there are different data cleaning techniques, in this study, cleaning was done using function structures with the Python programming language. With the preprocessing, the data was converted into the desired format. After data preprocessing, the word cloud of the data set was analyzed as positive and negative. The images of this examination are shown in Figure 3.

Fig. 3 Dataset word clouds

Positive labels, which are positive comments in the dataset, are replaced with 1 and negative labels, which are negative comments, are replaced with 0. Examples of comments and labels from the dataset for this preprocessing process are shown in Figure 4.

	review	sentiment
0	One of the other reviewers has mentioned that ...	positive
1	A wonderful little production. The...	positive
2	I thought this was a wonderful way to spend ti...	positive
3	Basically there's a family where a little boy ...	negative
4	Petter Mattei's "Love in the Time of Money" is...	positive
5	Probably my all-time favorite movie, a story o...	positive
6	I sure would like to see a resurrection of a u...	positive
7	This show was an amazing, fresh & innovative i...	negative
8	Encouraged by the positive comments about this...	negative
9	If you like original gut wrenching laughter yo...	positive

	review	sentiment
0	one of the other reviewers has mentioned that...	1
1	a wonderful little production the filming te...	1
2	i thought this was a wonderful way to spend t...	1
3	basically there s a family where a little boy...	0
4	petter mattei s love in the time of money i...	1
5	probably my all time favorite movie a story ...	1
6	i sure would like to see a resurrection of a ...	1
7	this show was an amazing fresh innovative ...	0
8	encouraged by the positive comments about thi...	0
9	if you like original gut wrenching laughter y...	1

Fig. 4 Before and after data preprocessing

C. *Splitting the Dataset*

A dictionary was created by controlling the number of unique words in the dataset. Each unique word in the dictionary was assigned a unique dictionary value. With these dictionary mappings of the words, the reviews were converted into a list of integers. This list was then sent to the model and the dataset was divided into two parts: train and test. The lexicon was created from the training data set only. The test data needs to be specific to the model. Because we do not know what kind of data will come in the future, it is very important that the test data has words unknown to the algorithm.

The first 40,000 reviews are allocated to the training data set and the remaining 10,000 reviews are allocated to the test data set. The same parts are used in all models to make it reasonable to compare the results. To count the unique words, we used the counter method, which is a subclass of dictionary in Python. This resulted in a total of 92,279 unique words in the training dataset.

D. *Digitizing the Dataset*

In order to classify the data set, it is necessary to digitize the data set. After the word dictionary was created, the reviews were converted into numerical form by replacing the integer index values corresponding to each word. Words that did not appear in the dictionary were assigned a common index. According to the word lengths of different reviews, the maximum word length of a review was set to 500. After these operations in the dataset, it was determined that the word lengths of the majority of the words analysed were close to 200. The histogram of this review is shown in Figure 5.

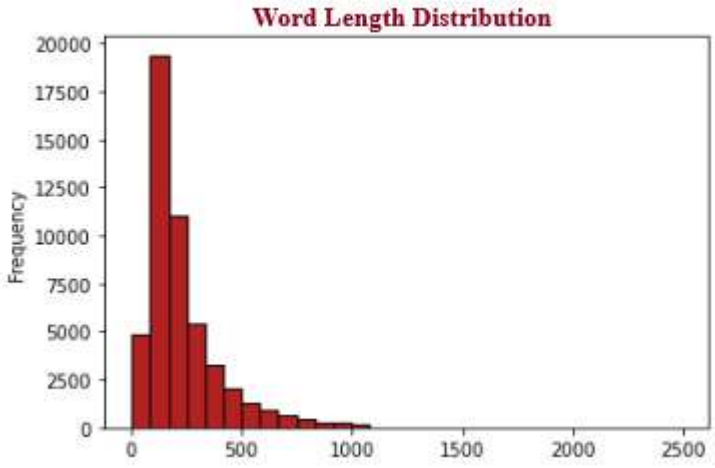

Fig. 5 Word length histogram

Once the labels have been converted into numerical forms so that positive emotions represent 1 and negative emotions represent 0, the dataset is ready for modelling.

E. Deep Learning

Deep learning is a subset of machine learning (ML) in which artificial neural networks (algorithms modelled to work like the human brain) learn from large amounts of data. Deep learning has recently been frequently preferred in sentiment analysis. In this study, MLP, LSTM, RNN, 1D CNN were used as deep learning algorithms.

F. Multi-Layer Perceptron (MLP)

MLP, which can also be thought of as a version of the logistic regression classifier, is a basic feed-forward neural network with at least one hidden layer. MLP is not a new neural network architecture, it has been widely used in machine learning applications for text classification problems even before the era of deep learning [19-22]. While MLP has the same number of input and output layers, it can also have more than one hidden layer. The MLP model developed in this study is built using the Tenserflow library of the Python programming language. ReLU and sigmoid functions are used as activation functions. An example of the MLP applying the appropriate activation functions to classify positive and negative comments is given in Figure 6.

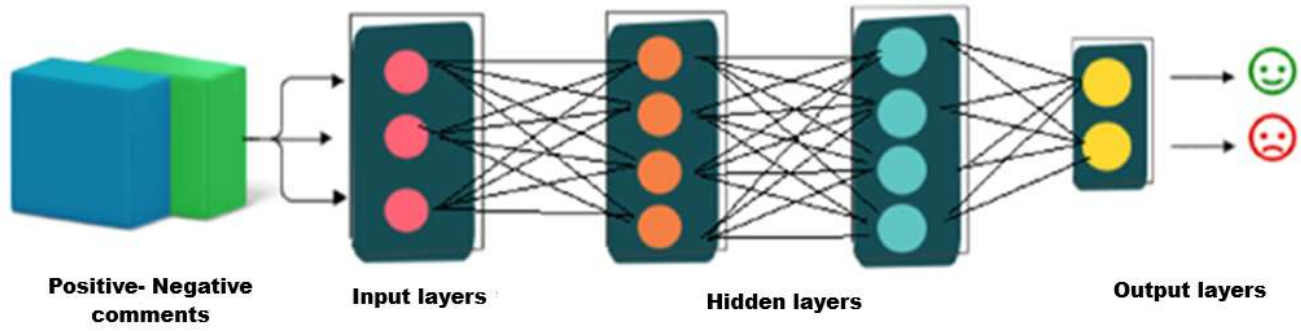

Fig. 6 MLP example

The sigmoid activated output layer generates a probability of sensitivity for each inspection. If this probability is close to 0, the sensitivity is negative, and if it is close to 1, the sensitivity is positive. Since this is a binary classification problem, the loss is compiled as 'binary_crossentropy'. The model consisting of 576,257 parameters was trained with 10 epochs and Adam optimizer and its performance was found to be 87.86% on the test dataset.

G. LSTM (Long Short Term Memory)

Long short-term memory networks are not very different from RNNs. However, LSTMs have memory cells for calculating hidden states. These cells are used to decide which data should be deleted or saved. In this study, an LSTM model with 353,365 parameters was designed and trained with 10 epochs and Adam optimizer and its performance was 87.77%.

H. RNN (Recurrent Neural Network)

Thanks to recent advances in technology, RNN can be used easily. RNN is a neural network model for learning existing patterns by exploiting sequential information [23,24]. RNN-based models are not good at capturing long-term contexts. In this model, it was concluded that they do not perform very well, achieving an accuracy of 82.80%.

İ. 1D CNN (Convolutional Neural Network)

Convolutional Neural Networks (CNN), also known as Convolutional Neural Networks (CNN), are faster to train than LSTM and RNN models. An accuracy of 91.43% was achieved on the test dataset. Although it has 354,755 fewer parameters, it has a better success than other models.

III. RESULTS

As can be seen from the results, the model with the best accuracy percentage is the 1D CNN architecture. It is followed by MLP and LSTM. The worst accuracy percentage belongs to the RNN architecture with approximately 83 percent. The data on the comparison of the models applied in the study are given in Table-2.

Table 2. Results of the models run with the data set

RESULTS FOR EMOTION CLASSIFICATION				
Model type	**Trainable parameters**	**Epoch Count**	**Optimizer**	**Accuracy (%)**
MLP	576.257	10	Adam	87,86
LSTM	353.365	10	Adam	87,77
1D CNN	354.755	10	Adam	91,43
RNN	328.465	10	Adam	82,80

IV. DISCUSSION AND CONCLUSION

This study found that the 1D CNN model outperformed the RNN models in the sentiment classification task performed on the IMDb dataset. This finding is important for evaluating the effectiveness of different architectures and structures of deep learning models on specific text processing tasks. In particular, the 1D CNN model may be more successful in tasks such as sentiment analysis, as it has the ability to more effectively identify structures in the horizontal plane in text data. These results contribute to the current debate in the literature on the effectiveness of different model architectures in text data processing.

As an example of an application that uses deep learning for emotion classification, the performance of the IMDb dataset on the models was compared. As a result of this comparison, it was observed that the 1D CNN model gave the best performance. It was determined that it gave a much better result than the RNN model. The IMDb dataset was made with 10 epochs in all models, that is, with the number of cycles (iterations), and they were optimized with Adam optimizer in the same way.

The results of this study provide an important contribution to evaluate the performance of deep learning models on sensitive and complex tasks such as sentiment classification. In particular, the ability of the 1D CNN model to more efficiently identify structures in the horizontal plane enables more accurate understanding and classification of emotional expressions in text data. This leads to higher accuracy and reliability in tasks such as sentiment analysis, enabling more effective results in real-world applications. Furthermore, the results of this study emphasize the importance of deep learning models in selecting the

most appropriate architecture for specific datasets and tasks. In particular, for sensitive tasks such as sentiment classification, model selection plays a critical role in determining the reliability and overall performance of the results. Therefore, in future research, the generalizability and applicability of the results of tests on different datasets and tasks should be evaluated from a broader perspective. In this way, a solid foundation can be established for a more effective and powerful use of deep learning models in text processing.

REFERENCES

[1] Bengio, Y. (2009). Learning deep architectures for AI. Foundations and Trends in Machine Learning 2(1), 1-127.

[2] Ravì, D., Wong, C., Deligianni, F., Berthelot, M., Andreu-Perez, J., Lo, B., & Yang, G. Z. (2017). "Deep learning for health informatics". IEEE journal of Biomedical and Health Informatics 21(1), 4-21.

[3] Lee, D. H. (2013). "Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks". Workshop on Challenges in Representation Learning, ICML (3), 2

[4] Cho, Y., & Saul, L. K. (2009). "Kernel methods for deep learning". Advances in Neural Information Processing Systems, 342-350.

[5] Lecun, Y., Bengio, Y., & Hinton, G. (2015). "Deep learning". Nature 521(7553), 436.

[6] Collobert, R., & Weston, J. (2008). "A unified architecture for natural language processing: Deep neural networks with multitask learning". International Conference on Machine Learning (ICML), 160-167

[7] Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., & Ng, A. Y. (2011). "Multimodal deep learning". International Conference on Machine Learning, 689-696.

[8] Deng, L., & Yu, D. (2014). "Deep learning: methods and applications". Foundations and Trends in Signal Processing, 7(3–4), 197-387

[9] Haque, M.R., Lima, S.A., & Mishu, S.Z. (2019). Performance Analysis of Different Neural Networks for Sentiment Analysis on IMDb Movie Reviews. In 2019 3rd International Conference on Electrical, Computer & Telecommunication Engineering (ICECTE). IEEE.

[10] Rao, G., et al. (2018). LSTM with sentence representations for document-level sentiment classification. Neurocomputing, 308, 49-57.

[11] Islam, M.M., & Sultana, N. (2018). Comparative study on machine learning algorithms for sentiment classification. International Journal of Computer Applications, 182(21), 1-7.

[12] Narayanan, V., Arora, I., & Bhatia, A. (2013). Fast and accurate sentiment classification using an enhanced Naive Bayes model. In International Conference on Intelligent Data Engineering and Automated Learning. Springer.

[13] Huang, Y., Zhang, X., Liu, Z., & Chen, Y. (2018). A topic BiLSTM model for sentiment classification. In Proceedings of the 2nd International Conference on Innovation in Artificial Intelligence.

[14] Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up? Sentiment classification using machine learning techniques. arXiv preprint cs/0205070

[15] Matsumoto, S., Takamura, H., & Okumura, M. (2005). Sentiment classification using word sub-sequences and dependency subtrees. In Pacific-Asia Conference on Knowledge Discovery and Data Mining. Springer.

[16] Arzu, M. ve Aydoğan, M. (2023). Türkçe Duygu Sınıflandırma İçin Transformers Tabanlı Mimarilerin Karşılaştırılması Analizi. Bilgisayar Bilimi, IDAP-2023 : Uluslararası Yapay Zeka ve Veri İşleme Sempozyumu(IDAP-2023), 1-6. https://doi.org/10.53070/bbd.1350405

[17] Kaggle, Url: https://www.kaggle.com/datasets/lakshmi25npathi/imdb-dataset-of-50k-movie-reviews

[18] Uçkan, T., Cengiz, H. A. R. K., Seyyarer, E., & Karcı, A. Ağırlıklandırılmış Çizgelerde Tf-Idf ve Eigen Ayrışımı Kullanarak Metin Sınıflandırma. Bitlis Eren Üniversitesi Fen Bilimleri Dergisi, 8(4), 1349 1362

[19] M.W. Gardner, S.R. Dorling, 1998, Artificial neural networks (the multilayer perceptron) a review of applications in the atmospheric sciences, Atmos. Environ. 32 2627–2636

[20] L. Deng, 2014, A tutorial survey of architectures, algorithms, and applications for deep learning, APSIPA Trans. Signal Inf. Process. 3.

[21] Santur, Y. (2020). Derin Öğrenme ve Aşağı Örnekleme Yaklaşımları Kullanılarak Duygu Sınıflandırma Performansının İyileştirilmesi. Fırat Üniversitesi Mühendislik Bilimleri Dergisi, 32(2), 561-570. https://doi.org/10.35234/fumbd.759131

[22] Santur, Y. (2019). Sentiment Analysis Based on Gated Recurrent Unit. In 2019 International Artificial Intelligence and Data Processing Symposium (IDAP) (pp. 1-5). Malatya, Turkey. doi: 10.1109/IDAP.2019.8875985

[23] M. M. Saritas, A. Yasar, "Performance Analysis of ANN and Naive Bayes Classification Algorithm for Data Classification," International Journal of Intelligent Systems and Applications in Engineering, vol. 7, pp. 88-91,2019, doi: 10.18201//ijisae.2019252786.

[24] S. Qing, H. Wenjie and X. Wenfang, "Robust Support Vector Machine with Bullet Hole Image Classification," IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), vol. 32, no. 4, pp. 440-448, 2002, doi: 10.1109/TSMCC.2002.807277.