Original Research Article

AE-FT: Auto Encoder Fixed-Target Training Features Extraction Approach for Binary Classification Problems

ABSTRACT

The main issues with machine learning-based feature extraction techniques are the requirement of extensive domain-level knowledge, experience, and the need to be supported by large amounts of data that are sometimes not available. Moreover, it is often difficult to apply domain-level knowledge to extract the necessary features for building a machine-learning classifier. To address these issues for binary classification problems, a novel feature extraction approach, *AE-FT(Fixed Target)* for extracting common features using a Deep Belief Network (DBN)-based Autoencoder (AE) is proposed in this paper. In this approach, common features are extracted by a Deep Belief Network (DBN) trained on a dataset sample's binary using *the Fixed Target training approach.*

The proposed common features extraction approach is tested and evaluated on two different data sets. For each dataset, the extracted features are used to train seven of the common machine learning binary classification algorithms and compared their performances. Moreover, the number of extracted features is very small compared to other existing feature extraction methods. Therefore, the proposed common features extraction method improves the performance of the binary classification algorithms by reducing the number of features reducing laborious processes and increasing the recognition accuracy effectively.

The results show that the proposed common features extraction approach, without any domain-level knowledge or human expertise, provides a very good performance compared to other feature extraction techniques.

Keywords: Features Extraction; Deep Learning; AE; Fixed-Target Training; Common Features.

1. INTRODUCTION

With the rapid development of machine learning technology, as a binary classification problem that helps people to find the law from the massive data to achieve the prediction effectively, data prediction has become an important part of people's daily lives.

Feature extraction is about finding a good data representation, which is very domain-specific, often requires human expertise, and is related to available measurements. The primary idea behind feature extraction is to compress the data with the goal of maintaining most of the relevant information. As to feature selection techniques, these techniques are

also used for reducing the number of features from the original feature set to reduce model complexity, and model overfitting, enhance model computation efficiency, and reduce generalization error. Therefore, it is better to better improve the accuracy of the learning algorithm and shorten the training and output time.

The feature extraction methods are useful for different applications as mentioned in [1], such as social science, healthcare, environment, agriculture, spam filtering, antivirus technology, economics, medical diagnosis, face recognition, action recognition, speech recognition, gesture recognition, marketing, wireless network, gene expression, software fault detection, internet traffic prediction, etc. Therefore, the research of machine learning algorithms in feature extraction problems is a research hotspot in recent years. There are two broad categories for feature extraction algorithms: linear and nonlinear.

Linear feature extraction assumes that the data lies in a linear subspace. Use matrix factorization to protect them. On the other hand, in nonlinear feature extraction or dimensionality reduction, a low-dimensional surface can be mapped into a high-dimensional space so that a nonlinear relationship among the features can be found and easily detected. Theoretically, a transformation function f(x) can be used to map the features into a higher-dimensional space and then mapped back into the lower-dimensional space, so that the relationship can be viewed as nonlinear.

1.1. Kernel PCA (KPCA)

Kernel PCA introduced by Scholkopf et al. [2], is an extension of PCA that allows for the separability of nonlinear data by making use of kernels. The basic idea behind it is to project the linearly inseparable data onto a higher dimensional space where it becomes linearly separable. Unfortunately, it has a serious limitation in terms of space complexity since it stores all dot products of the training set and therefore the size of the matrix increases quadratically with the number of data points as presented in [3].

Another drawback of the KPCA, however, is the cost of computation could be extremely high, which could lead to the attendant numerical problems of diagonalizing large matrices, which limits its applicability in many large dataset problems. But, an Expectation-Maximization (EM) algorithm for KPCA to overcome these drawbacks was proposed in [4], which is an expectation-maximization approach for performing kernel principal component analysis. Experimental results showed that EM is an efficient method computationally, especially for a large number of data points..

1.2. Locally Linear Embedding (LLE)

Locally Linear Embedding, proposed by Saul et al. [5], is a dimensionality reduction technique based on Manifold Learning that involves the computation of low-dimensional neighborhood preserving embeddings of inputs that are of high dimension in nature. Manifold Learning aims to make a manifold object, an object of D dimensions that is embedded in a higher-dimensional space, representable in its original D dimension instead of being represented in an unnecessarily greater space.

LLE has the ability to learn the global structure of nonlinear manifolds like those from images of faces or documents of text by exploiting the local symmetries of linear reconstructions. LLE has been applied successfully in a wide range of applications which includes face recognition and remote sensing, MRI, shape analysis of the hippocampus in AD, diffusion tensor imaging, breast lesion segmentation, feature fusion, and image classification according to [6].

LLE is popular among researchers because of its ability to deal with large data sets of high-dimensional data and its non-iterative way of finding embeddings. However, it has some drawbacks which include sensitivity to noise, the inability to deal with novel data, and the inevitable ill-conditioned Eigen problems. Some efforts have recently been made to develop extensions of the classical LLE.

Supervised and semi-supervised versions of LLE were proposed by [7] and [8], respectively, for plant classification based on images of leaves.

1.3. Linear Discriminant Analysis (LDA)

LDA is a supervised learning dimensionality reduction, feature extraction technique, and Machine Learning classifier that was invented by Fisher et al. [9]. LDA uses within-classes and between-classes measures by maximizing the distance between the mean of each class and minimizing the spreading within the class itself. This is a good choice because maximizing the distance between the means of each class when projecting the data in a lower-dimensional space can lead to better classification results.

An advantage of LDA is that it is able to use information from both features to create a new axis which in turn minimizes the variance and maximizes the class distance of the variables.

Although the LDA is one of the most well-used data reduction techniques, it has a number of limitations. The small sample problem (SSS), is one of the main problems of LDA, which happens when the dimensions are much higher than the number of samples in the data matrix, LDA is unable to find the lower dimensional space resulting in the within-class matrix becoming singular. Different approaches have been proposed to solve this problem, such as what was proposed in [10] and [11]. In addition to the assumption that the input data follows a Gaussian Distribution, therefore applying LDA to not Gaussian data can lead to poor classification results.

A semisupervised variant of LDA, which performed better than the classical LDA, was proposed by [12] that mainly combines both labeled and unlabeled data for training LDA and allows using LDA for the situation where the labeled data are few.

Application of LDA includes facial recognition, text recognition, automatic diagnosis of machine operations, early detection of diseases, person reidentification, hand movement classification, motor imagery EEG, and groundwater redox conditions.

1.4. t-distributed Stochastic Neighbor Embedding (t-SNE)

t-Stochastic Neighbor Embedding (t-SNE) is an unsupervised Non-linear Dimension Reduction Technique (NLDRT) that was introduced by Maaten et al. [13]. The technique is a variation of the Stochastic Neighbor Embedding introduced by Hinton et al.[8], whose main objective is the construction of probability distributions from pairwise distances such that larger distances correspond to smaller probabilities and vice versa. t-SNE is typically used to visualize high-dimensional datasets, it works by minimizing the divergence between a distribution constituted by the pairwise probability similarities of the input features in the original high-dimensional space, which is modeled using a Gaussian Distribution and its equivalent in the reduced low-dimensional space, modeled using a Student's t-distribution.

t-SNE makes use of the Kullback-Leiber (KL) divergence in order to measure the dissimilarity of the two different distributions, as mentioned in [14]. The KL divergence is then minimized using gradient descent.

t-SNE is the most commonly used in single-cell analysis. However, it has some limitations as mentioned in [15]. The limitations include slow computation time, the inability to meaningfully represent very large datasets, and the loss of large-scale information.

1.5. Deep Learning Approach

The major difference between deep learning and traditional pattern recognition methods is that deep learning automatically learns features from big data, instead of adopting handcrafted features, as stated in [16]. Deep learning is able to quickly acquire new effective feature representations from training data.

Autoencoders

Autoencoders are feed-forward neural networks that were first introduced by Rumelhart et al. [17]. They can learn a compressed and distributed representation of data, which can be used as a dimensionality reduction or feature extraction technique. They use nonlinear transformations to project data from a high dimension to a lower one. An autoencoder usually has at least one hidden layer between the input and output layers. The number of neurons in the hidden layer is usually set to less than those in the input and output layers, thus creating a bottleneck, with the intention of forcing the network to learn a higher-level representation of the input as presented in [18].

Autoencoders are typically trained in an unsupervised manner, using backpropagation with stochastic gradient descent, to approximate a function by which data can be classified, as mentioned in [19] and [20]. For every training input, the difference between the input and the output is measured (using squared error) and it is back-propagated through the neural network to perform weight updates on the different layers.

Compared with other machine learning methods, deep learning is able to detect complicated interactions in features, learning lower-level features from nearly unprocessed original determine characteristics that are not easy to be detected. Furthermore, they hand class members with high cardinal numbers and process untapped data. Unfortunately, if all

input features are independent of each other, then the autoencoder will find it particularly difficult to encode the input data into a lower-dimensional space.

The advantages are higher discriminating power and control overfitting when it is unsupervised, and the disadvantages are loss of data interpretability and transformation may be expensive.

There are many other deep learning-based feature extraction approaches. A skeleton-based abnormal gait recognition approach was proposed by Jun et al. [21]. They proposed a feature extraction method using the RNN AEs to minimize the irrelevant information of the original skeleton data. They used two-step training of a hybrid RNN and AE-DM model and approved that it is more effective than the single-step training of the End-to-End model that has the same data flow. Ma and Yuan [22] proposed a method for extracting features from images based on deep CNN and PCA. They used a neural network to extract features and a PCA algorithm for feature dimension reduction. Then they compared the performance of the PCA before and after the improvement claim achieving memory, and time optimization. Moreover, the SVM classifier accuracy was enhanced. However, these are domain-dependent approaches that work only for the specific domain not for binary classification problems in general.

Table 1. Comparison of Feature Extraction Methods.

Feature Extraction Technique	Domain level Knowledge	Data preprocessing	Limitations
Kernel PCA	yes	yes	space complexity
LLE	yes	yes	sensitivity to noise,the inability to deal with novel data and the inevitable ill-conditioned Eigen problems
LDA	yes	yes	small sample problem (SSS)
t-SNE	yes	yes	slow computation time
AE	often	yes	Loss of data interpretability Transformation may be expensive

The main issue with machine learning-based feature extraction techniques is the requirement of extensive domain-level knowledge and experience, see Table 1. Moreover, it is often difficult to apply domain-level knowledge to extract the necessary features for building a machine-learning classifier. Therefore, it is significantly important to find and develop feature extraction techniques that depend mainly on the training data and don't require or depend on domain-level knowledge and experience, and this is our main purpose.

This paper presents the use of denoising stacked autoencoders with supervised fixed-target training in order to extract the common features of the training data that can be used in binary classification. The proposed approach achieves 73.50% accuracy, which is so far a good result that does not need any domain expertise.

In the experiments, we used the extracted common features to build binary classifiers using seven binary classification methods Naive Bayes, Logistic Regression, K-Nearest Neighbors, Support Vector Machine, Decision Tree, Random Forest, and Voting Classification- for binary classification.

The proposed method, fixed-target training of a deep stacked autoencoder, enabled a good recognition accuracy, better generalization, and more stability than that which could be achieved with the other methods. In contrast to most existing approaches that normally have a separate stage for data preprocessing, followed by domain-dependent feature extraction. We developed a domain-independent deep neural network framework for common feature extraction which enables us to easily, without the need for domain-level knowledge or expertise, extract features that can be effectively used in binary classification problems.

2. METHODOLOGY

In this section, we describe our proposed novel deep learning-based approach for common feature extraction in detail. The main question we are trying to answer is the following:

Is it possible to extract the common features from the raw binary representations without any domain expertise of a given dataset that could be used in binary classification?

In recent years, deep learning methods have proven very successful in accomplishing dimensionality reduction and feature extraction tasks in many domains, especially computer vision and cybersecurity according to [24, 25, 26, 27]. The proposed methodology works as follows:

Firstly, training the proposed common features extraction model to extract features to use in binary classification using various common machine learning classification algorithms.

Secondly, select seven binary classification algorithms and use the extracted common features to build binary classifiers using common algorithms.

Our method uses stacked denoising autoencoders for extracting the common features of the training dataset that can be used effectively in binary classification. The input to the Deep Belief Network (DBN) has a fixed length, but the dataset sample length is variable. In order to represent the dataset binary as a fixed-sized vector, which would be the input to the neural network, we repeatedly pad the sample binary until the specified size is met. This process gives better results than 0's or 1's padding that the DBN learns as a common feature. Then we use two different datasets, the IMDB dataset [28] and the Enron-Spam dataset [29], to train and test the proposed feature extractor for extracting the common features and then used for training binary classification models using the extracted features of the binary representation of the datasets. We focus on the top seven most common binary classification algorithms Naive Bayes, Logistic Regression, K-Nearest Neighbours, Support Vector Machine, Decision Tree, Random Forest, and Voting Classification.

2.1. Deep Belief Network Fixed-Target Training

In stacked denoising autoencoders, first introduced by Vincent et al. [23], the data at the input layer is replaced by noised data while the data at the output layer stays the same; therefore, the autoencoder can be trained with much more generalization power, according to [16].

Usually, denoising autoencoder training is unsupervised, the input sample is corrupted by adding noise (or more often by zeroing the values). That is, given an input **S**, first it is corrupted to **S**^ and then fed to the input layer of the network. The objective function of the network in the output layer remains to generate the uncorrupted version of the input (see Figure 1-(a)). But in our approach, we use a novel supervised training strategy, which we call a *fixed-target training* strategy.

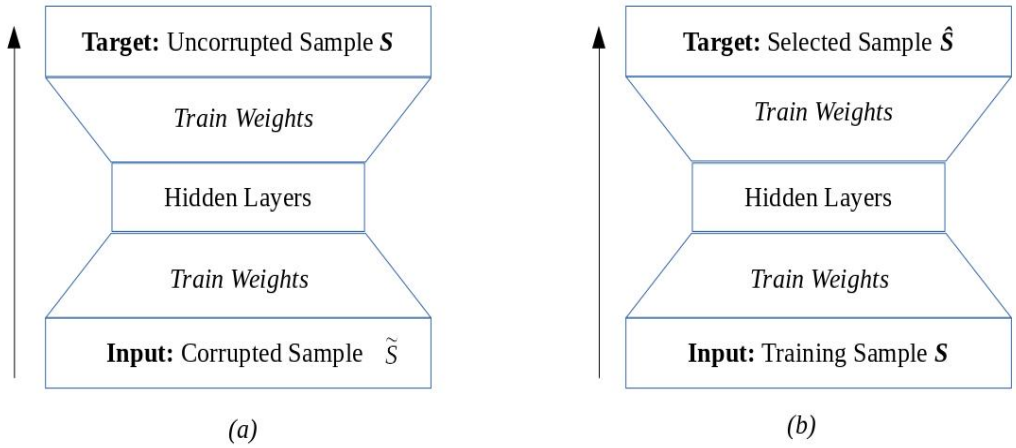

Fig. 1. Denoising autoencoder training. (a)-traditional unsupervised training, (b)-fixed-target supervised training.

In the *fixed-target training* strategy, we randomly select one of the training samples **S**^, fix it in the output layer, and for every input **S**^ of the training samples, the objective function of the network is to generate **S**^, (i.e. we consider all training samples as corrupted versions of the selected sample **S**^ (see Figure 1-(b)). So, the "hidden units" of DBN compute internal representations analogous to the extracted common features.

This training approach works better than traditional training. By fixing the target output, the network is forced to generalize better and determine more high-level common patterns. Moreover, the network learns better even when few training samples are available.

When a DBN's training is complete, we can discard the decoder layer, fix the values of the encoder layer, and use the encoder as the common features extractor. In a typical implementation, the extracted features may then be used for supervised binary classification.

In order to achieve our goal, we create a deep belief network (DBN) by training a deep stack of denoising autoencoders. We use fixed-target training to train a deep denoising autoencoder consisting of five layers: 8,192–2,048–512–128–32. At the end of this training phase, we have a deep network that is capable of converting 8,192 length input vectors into 32 floating point feature values. Note that the network is trained only using the samples in the training set, and for all future samples it will be run in prediction mode;(i.e., receiving the

8,192-sized vector it will produce 32 output values, without modifying the weights). See Figure 2.

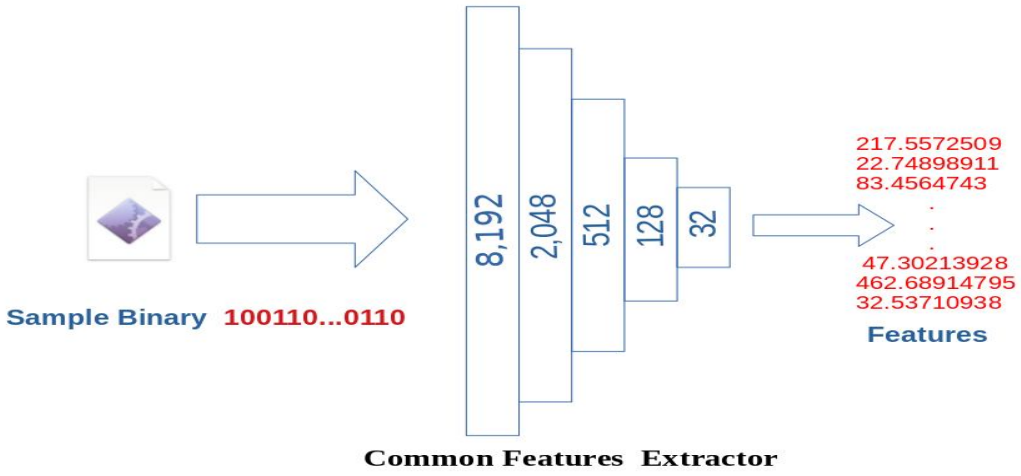

Fig. 2. Illustration of the common features extraction stages from feeding the sample binary to features extraction using DBN.

The sample's binary bit string is fed to the neural network, and the deep neural network generates a 32-sized vector at its output layer, which we treat as the common feature values of the sample.

The next section provides the implementation details and experimental results and demonstrates that the resulting 32-sized vector (i.e., the extracted common features) indeed provides a good representation of the data common feature.

2.2. Datasets

We are using two different datasets, the Internet Movie Database (IMDB) dataset [28] and the Enron-Spam dataset [29], for testing our approach. The IMDB dataset, Large Movie Review Dataset v1.0, of highly polar movie reviews in the form of text comments on different movies and a positive or negative score. This dataset contains movie reviews along with their associated binary sentiment polarity labels. It is intended to serve as a benchmark for sentiment classification. The core dataset contains 50,000 reviews split evenly into 25k train and 25k test sets. The overall distribution of labels is balanced (25k pos and 25k neg). Furthermore, the train and test sets contain a disjoint set of movies, so no significant performance is obtained by memorizing the movie's unique terms and they are associated with the observed labels. In the labeled train/test sets, a negative review has a score < 4 out of 10, and a positive review has a score > 7 out of 10. Thus, reviews with more neutral ratings are not included in the train/test sets.

These datasets have been used as a benchmark in several research papers on spam filtering, text classification, and natural language processing [30, 31]. Therefore, the results of this work are hopefully comparable to other similar works within the area, without having to account for unique datasets.

The goal is to train the proposed feature extractor for extracting the common features and then used it for training binary classification models using the extracted features of the binary representation of the datasets. We focus on the top seven most common binary classification algorithms Naive Bayes, Logistic Regression, K-Nearest Neighbours, Support Vector Machine, Decision Tree, Random Forest, and Voting Classification.

2.3. *Common Features Extraction Method*

As described in the previous section, we use a fixed-target training strategy to train a deep denoising autoencoder consisting of five layers (8,192–1,024–512–128–32). We randomly pick one of the training samples as a pivot sample, and each time a new input is given to the network we put the pivot sample in the output as the target. This way, we enforce the autoencoder to learn the common features of the training samples.

Fixed-target training is essential for the network to learn common features in a very small time and even when few samples are available. We use rectified linear units (ReLU) for the nonlinearity function when training deep neural networks to diminish the gradient vanishing problem and result in faster convergence, as approved in [12, 18].

We build the model using machine learning algorithms in *the Keras library*. Other parameters we use are 50 training epochs, a learning rate of 0.003, a batch size of 10, and no momentum.

Figure 3 shows the learning curves, which approve the feasibility of our proposed common feature extraction approach.

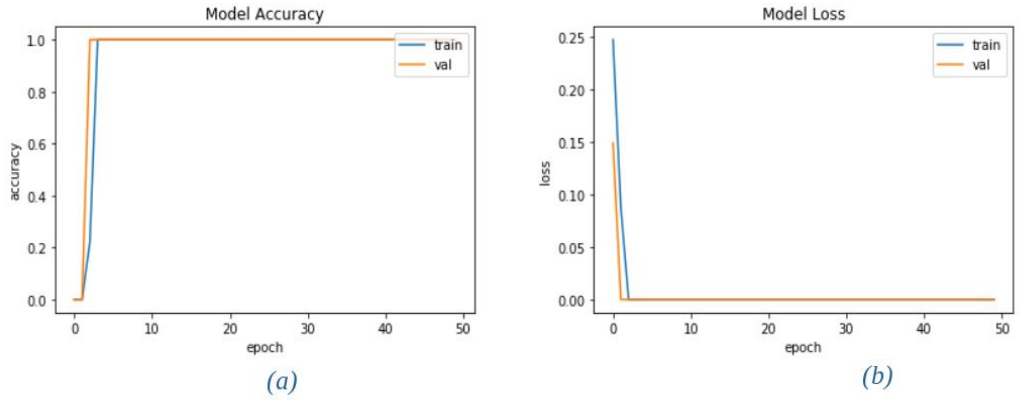

Fig. 3. Common Features Extraction Model training and validation. (a)-Accuracy, (b)-Loss

3. EXPERIMENTAL RESULTS

In this section, we conducted various experiments to evaluate the features of the DBN AEs for binary classification. First, we compared the number of features used in [28,

29] and the extracted common features. Then, we conducted experiments to show the effectiveness of the proposed DBN AEs. Finally, we showed the performance of the features of the DBN AEs used in binary classification problems, we used seven binary classification methods Naive Bayes, Logistic Regression, K-Nearest Neighbours, Support Vector Machine, Decision Tree, Random Forest, and Voting Classification- for binary classification.

In the experiments, we trained a deep denoising autoencoder consisting of five layers, 8192–2048–512–128–32, using a Fixed-Target training strategy. We used only positive training samples for training the common features extractor. Then, we used the common features extractor for extracting the values of the common features of the training samples and used it for training the binary classifiers.

In the testing phase, we first extract the values of the common features of the sample and pass them to the binary classifier for classification. See Figure 4.

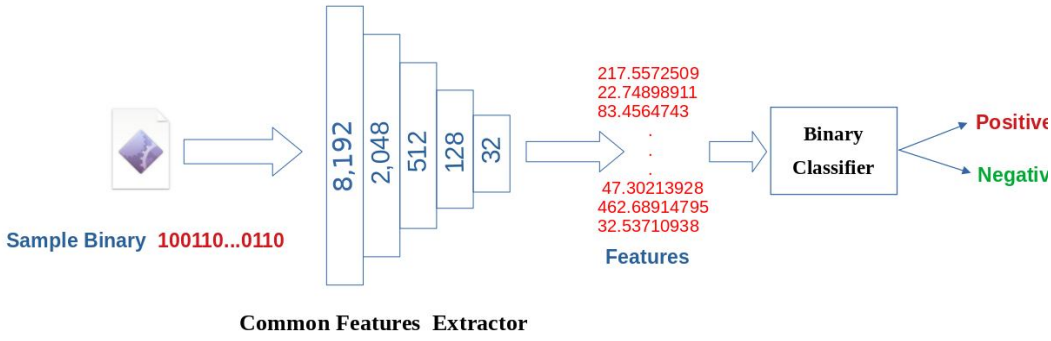

Fig. 4. Overview of our deep learning approach for common features extraction, illustration of all stages from feeding the sample binary to the Common Features Extractor ending with the decision made by the binary classifier.

The results demonstrate that, with proper structure and parameters, the performance of the proposed deep learning method on common feature extraction is useful even in the lack of domain expertise.

Table 2. Performance of Binary Classification Methods, trained by the common features on IMDB and Enron-Spam datasets.

Binary Classification Method	IMDB Dataset		Enron-Spam Dataset	
	Training Score	Test Score	Training Score	Test Score
Naive Bayes	63.0%	60.7%	63.0%	66%
Logistic Regression	88.0%	73.5%	87.0%	73%
K-Nearest Neighbours	87.3%	73.2%	88.7%	73.2%
Support Vector Machine	67.0%	66.3%	67.3%	67%

Decision Tree	100%	70.5%	100%	72%
Random Forest	93.5%	73.4%	93.5%	73.5%
Voting Classification	85.0%	72.7%	84.7%	73%

The accuracy results are given in Table 2, test samples are 73.5% correctly classified by conducting a Logistic Regression classifier for the proposed approach with the extracted common features of the IMDB dataset. For the Enron-Spam dataset, the best results are obtained by Random Forest 73.5%. The binary classification performances of the approaches are given in Table 2. According to the results, the common features extracted by the proposed approach give a very good performance for all seven binary classification methods. Logistic Regression and Random Forest are the most effective classifiers for this approach.

Compared with other methods, the proposed approach fixed-target training of a denoising deep stacked autoencoder achieves 73.50% accuracy, which is very good for binary classification without any need for domain-level expertise or data preprocessing.

4. DISCUSSION AND EVALUATION

We trained and tested, using two different datasets, the proposed feature extractor for extracting the common features. Then, we used the extracted features for training seven of the most common binary classification algorithms Naive Bayes, Logistic Regression, K-Nearest Neighbours, Support Vector Machine, Decision Tree, Random Forest, and Voting Classification.

According to the results reported in Table 2, the best accuracy values are obtained by conducting Logistic Regression and Random Forest classifiers. When comparing the number of features used, 32 common features were extracted using the proposed feature extraction algorithm, on the other hand, 3000 and 10000 features were used in [28] and [29], respectively, in addition to the preprocessing carried to extract the features used in [28] and [29]. However, by applying the proposed common features extraction algorithm, it is observed that high success rates are achieved with very few features and increase the overall process performance.

5. FUTURE WORK

Future work could include examining how to improve the accuracy of the proposed common feature extraction method by finding the optimal autoencoder structure and activation function. Furthermore, examining different ways for the pivot sample selection. Of course, we could also look at different types of classifiers and use different datasets in different domains. One could also see is it useful to use the proposed method in data collection; when only a few samples are available.

Finally, it would be interesting to see if the proposed method could mitigate the problem of the need for domain expertise and data preprocessing for feature extraction and the need for a big dataset to train binary classifiers.

6. CONCLUSION

In binary classification, feature extraction is a basic and important matter because the original data contain noise and irrelevant information which decreases the classification accuracy. to reduce the dimensionality of the feature vector space.

In this paper, we propose a feature extraction method using the DBN. The proposed approach, AE fixed-target supervised training, extracts the common characteristics of the original data and minimizes the irrelevant information without the need for any domain-level expert or expensive data preprocessing. Furthermore, it needs less training data and produces fewer features compared to other feature extraction techniques. Therefore, the extracted features improve the binary classification performance.

The most interesting result of our evaluation was the very good performance of our common features extension approach using all binary classification algorithms tested on both datasets that have been used, but further experiments are needed to be more confident.

Data Availability: Tow Datasets are used, The IMDB and Enron-Spam data supporting this research are from previously reported studies and datasets, which have been cited. The processed data are available at http://www.cs.cornell.edu/people/pabo/movie-review-data/ and http://www2.aueb.gr/users/ion/data/enron-spam/ , respectively.

ACKNOWLEDGEMENTS

Funding: The study attracted no funding.

Conflict of Interest: All authors declare that they have no conflict of interest.

REFERENCES

1. Ghojogh, Benyamin, Maria N. Samad, Sayema Asif Mashhadi, Tania Kapoor, Wahab Ali, Fakhri Karray, and Mark Crowley. Feature selection and feature extraction in pattern analysis: A literature review. *arXiv preprint arXiv:1905.02845* (2019).
2. Schölkopf, Bernhard, Alexander Smola, and Klaus-Robert Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural computation* 10, no. 5 (1998): 1299-1319.
3. Liu, Xiaofang, and Chun Yang. Greedy kernel PCA for training data reduction and nonlinear feature extraction in classification. In *MIPPR 2009: Automatic Target Recognition and Image Analysis*, vol. 7495, pp. 768-775. SPIE, 2009.
4. Rosipal, Roman, and Mark Girolami. An expectation-maximization approach to nonlinear component analysis. *Neural Computation* 13, no. 3 (2001): 505-510.
5. Saul, Lawrence K., and Sam T. Roweis. Think globally, fit locally: unsupervised learning of low dimensional manifolds. *Journal of machine learning research* 4, no. Jun (2003): 119-155.
6. Nanga, Salifu, Ahmed Tijani Bawah, Benjamin Ansah Acquaye, Mac-Issaka Billa, Francis Delali Baeta, Nii Afotey Odai, Samuel Kwaku Obeng, and Ampem Darko Nsiah. Review of Dimension Reduction Methods. *Journal of Data Analysis and Information Processing* 9, no. 3 (2021): 189-231.

7. Kouropteva, Olga, Oleg Okun, and Matti Pietikäinen. Incremental locally linear embedding algorithm. In *Scandinavian Conference on Image Analysis*, pp. 521-530. Springer, Berlin, Heidelberg, 2005.
8. Pan, Yaozhang, Shuzhi Sam Ge, and Abdullah Al Mamun. Weighted locally linear embedding for dimension reduction. *Pattern Recognition* 42, no. 5 (2009): 798-811.
9. Xanthopoulos, Petros, Panos M. Pardalos, and Theodore B. Trafalis. Linear discriminant analysis. In *Robust data mining*, pp. 27-33. Springer, New York, NY, 2013.
10. Sharma, Alok, and Kuldip K. Paliwal. A new perspective to null linear discriminant analysis method and its fast implementation using random matrix multiplication with scatter matrices. *Pattern Recognition* 45, no. 6 (2012): 2205-2213.
11. Belhumeur, Peter N., Joao P. Hespanha, and David J. Kriegman. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. IEEE Transactions on pattern analysis and machine intelligence 19, no. 7 (1997): 711-720.
12. Zhang, Yu, and Dit-Yan Yeung. Semisupervised generalized discriminant analysis. *IEEE Transactions on Neural Networks* 22, no. 8 (2011): 1207-1217.
13. Van der Maaten, Laurens, and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of machine learning research* 9, no. 11 (2008).
14. Hinton, Geoffrey E., and Sam Roweis. Stochastic neighbor embedding. *Advances in neural information processing systems* 15 (2002).
15. Xie, Bo, Yang Mu, Dacheng Tao, and Kaiqi Huang. m-SNE: Multiview stochastic neighbor embedding. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* 41, no. 4 (2011): 1088-1096.
16. Wang, Haohan, and Bhiksha Raj. On the origin of deep learning. *arXiv preprint arXiv:1702.07800* (2017).
17. Rumelhart, David E., Geoffrey E. Hinton, and Ronald J. Williams. *Learning internal representations by error propagation*. California Univ San Diego La Jolla Inst for Cognitive Science, 1985.
18. David, Omid E., and Nathan S. Netanyahu. Deepsign: Deep learning for automatic malware signature generation and classification. In *2015 International Joint Conference on Neural Networks (IJCNN)*, pp. 1-8. IEEE, 2015.
19. Werbos. Beyond Regression: New Tools for Prediction and Analysis in the Behavioral Science, Appl. Math. Harvard University (1974).
20. Hinton, Geoffrey E., and Ruslan R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *science* 313, no. 5786 (2006): 504-507.
21. Jun, Kooksung, Deok-Won Lee, Kyoobin Lee, Sanghyub Lee, and Mun Sang Kim. Feature extraction using an RNN autoencoder for skeleton-based abnormal gait recognition. *IEEE Access* 8 (2020): 19196-19207.
22. Ma, Ji, and Yuyu Yuan. Dimension reduction of image deep feature using PCA. *Journal of Visual Communication and Image Representation* 63 (2019): 102578.
23. Vincent, Pascal, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, Pierre-Antoine Manzagol, and Léon Bottou. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of machine learning research* 11, no. 12 (2010).
24. Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM* 60, no. 6 (2017): 84-90.
25. Lee, Honglak, Chaitanya Ekanadham, and Andrew Ng. Sparse deep belief net model for visual area V2. *Advances in neural information processing systems* 20 (2007).
26. Le, Quoc V. Building high-level features using large scale unsupervised learning. In *2013 IEEE international conference on acoustics, speech and signal processing*, pp. 8595-8598. IEEE, 2013.
27. Simonyan, Karen, and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).

28. Maas, Andrew, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, pp. 142-150. 2011.
29. Metsis, Vangelis, Ion Androutsopoulos, and Georgios Paliouras. Spam filtering with naive bayes-which naive bayes?. In *CEAS*, vol. 17, pp. 28-69. 2006.
30. Eyheramendy, Susana, David D. Lewis, and David Madigan. On the naive bayes model for text categorization. In *International workshop on artificial intelligence and statistics*, pp. 93-100. PMLR, 2003.
31. Gómez-Hidalgo, José María, Manuel J. Maña López, and Enrique Puertas Sanz. Combining text and heuristics for cost-sensitive spam filtering. In *Fourth Conference on Computational Natural Language Learning and the Second Learning Language in Logic Workshop*. 2000.