Explainability of One Class Classification Anomalous Text Detection Models

Kamryn Parker

April 28, 2022

Abstract

The field of deep learning and AI is ever-expanding for their algorithmic power and performance that traditional machine learning models are not able to provide. As the field expands for broader use, the power of a deep learning model may be enticing but, being able to understand how a model makes its decisions is an important and needed source of model validity. The Context Vector Data Description model (2019) is a one-class text anomaly detection algorithm that provides meaningful insights for anomaly text detection. The model provides room to be able to interpret the results using explainability techniques like LIME. This paper utilizes the models results to present a novel approach to the interpretability of the Context Vector Data Description model and why the decisions made are important to the overall model success. In the exploration we found that typical explanation models were not viable for this type of experiment. We test a different approach by surveying the attention mechanism of the model as a whole to examine its relationship to model interpretation. It was found that the while attention weights were said to be indicators of important anomalous texts, there was not enough significant evidence to overtly claim this idea.

1 Introduction

Deep learning models are becoming increasingly popular in machine learning research and industry level insights. Deep learning models are powerful models that can automate routine labor, understand speech or images, make diagnoses in medicine, and support basic scientific research [GBC17]. However, while these models may produce substantial results, there are some drawbacks to their commercial use among companies and research groups. The largest drawback is the difficulty in understanding the deep learning model's rationale for the decisions they make. The reason many deep learning models are difficult to explain are for their complex hidden layers whose decisions are not easily interpretable to model architects. Deep learning architecture is typically built so that it only has the input variables and output variables visible. The main process is considered to be a "black-box" of decision making, that is, their internal logic cannot be easily understood by a human [ATS20]. So, this can lead to a lot of mistrust with the learning model's capabilities if researchers and stakeholders are unable to explain how the model developed its results.

The research area of explainable modeling is important to deep learning and their use in modern day infrastructure. Explainabiltiy is the field of research that makes an effort to interpret a deep learning model and the features the models uses to make its decisions [ASLA20]. It is as equally important for deep learning researchers to have interpretable models as it is to building them in general. Through the process of explaining deep learning models, it may even become possible to improve these deep learning models overall to reach their highest capabilities [SWM17].

There are typically two types of explainability for models, transparency modeling and post-hoc analysis. Transparency modeling attempts to uncover what the model's inner functions are (i.e. model structure, training architectures, etc.). Post-hoc analysis works to explain the model after the training and testing stages. Typically post-hoc analysis includes visualizations or statistical statements of the choices the model made during the training and testing phase [XUD+19]. For the purposes of this paper the goal is to explain the choices a model makes in a post-hoc analysis of the classification tool.

Therefore, can an explainable model be developed for the model we will evaluate, the Context Vector Data Description anomalous text detection model? We hypothesize there is strong indication

that using a previously defined explainable model, can assist in accurately explaining the results of the Context Vector Data Description model in a post-hoc analysis. This paper outlines the Context Vector Data Description model and its dependencies that point to interpretation results.

The attempts in achieving an interpretable model was unsuccessful so, we instead adjusted our focus to analyzing the attention weights and context vectors of this model. We examined a comparison between results for a large attention weight scheme and a small attention weight scheme. There were significant differences between the two results which led us to call into question the validity of self attention weights as interpretation mechanisms for a model.

2 Related Works

2.1 Text Classification

Text classification is a type of classification algorithm that attempts to classify a document under a predefined category [IKT]. Text classification algorithms are powerful tools that can help users better define their data sets and text corpora by being able to classify various elements in each document type. Understanding text can also help us understand human behavior. Being able to identify which words are most significant in documents are key to understanding how text classification algorithms make decisions.

The model we will be analyzing uses anomaly detection techniques to classify anomalous texts. Anomaly detection (AD) is the practice of finding patterns in data that do not conform to the expected behavior [CBK09]. The advantages of using text classification with anomaly detection is the ability to map interdependencies between variables and predictions to better classify anomalous texts [TS18]. Text AD findings can integrate with both historical information and current data.

Text classification algorithms pose a large market of opportunity to make informed decisions about "non-numerical" data. We focus on text classification anomaly detection for this paper because while a powerful tool, it is also under researched, expressly in the field of One Class Classification.

2.2 One Class Classification

One class classification (OCC) modeling is used when the negative class is either absent, poorly sampled or not well defined [KM10]. This technique forces the model to learn in the constraints of the positive class only. Many times, this type of model is used for anomaly detection or concept learning as these research areas typically lack data from the negative class [HFW08]. Explainability modeling for one class classification models, specifically in anomaly detection is a growing area of research.

Combining one class classification with anomaly detection is an approach that is popular with anomaly image classification. This model type can help uncover important patterns in data, especially medical diagnostics like tumor recognition [WRH+18] and other types of image classification using only one class (positive or negative) which would be difficult to accomplish otherwise. A popular technique in this domain is to use One Class Support Vector Machines (OC-SVMs) for large scale anomaly detection because of three main advantages, (i) does not require an explicit statistical model, (ii) provides an optimum solution for classification by maximizing the margin of the decision boundary, and (iii) avoids the curse of dimensionality problem [ZMH09]. OC-SVMs have been extensively studied for their ease of interpretability.

We will concentrate on a less explored research area in OCC, text classification. Text classification has been found to be successful in anomaly one class classification on pre-trained word-embeddings [RZV+19]. The experimental model in Ruff et. al's paper will be the focus of this paper because of its capability to represent a learning approach for AD on text which has not been studied previously. However, a drawback of this model is that it is a fairly new approach to OCC models. Current research on explainability techniques in OCC have been found successful using image anomaly detection but have not been attempted for text anomaly detection for OCC [LRV+20]. This means there is not much research on this model's interpretability even though the CVDD model's architecture assures an ease of interpretability. Therefore there is space to explore the possibility of developing an interpretation model in this context.

3 Model and Approach

In Ruff et al's paper [RZV+19] on anomaly text classification the authors introduce Context Vector Data Description (CVDD) as a one class classification anomaly detection approach. The model builds upon existing word embedding models to learn different sentence representations which allow the ability to capture multiple semantic contexts via a self-attention mechanism. The CVDD self-attention mechanism is a deep learning model that utilizes the GloVe pre-trained word embeddings. Ruff et. al claims that using this model and mechanism will provide for ease of interpretability for contextual anomaly model detection. The CVDD paper discusses the different properties of the model, including its supposed inherent interpretability, and its results but does not specify the reasoning behind decision features the model uses to make predictions. They chose to train their word-embeddings with context vectors to jointly allow the algorithm to capture multiple modes of normalcy, which they allege will enhance model interpretability. Considering the CVDD model paper does not explore, in depth, the attention weight mechanism's innate interpretability, we are allowed to ask questions about the model's explainability and attempt to tackle how the model makes informed decisions about one class text data.

3.1 Local Interpretable Model-agnostic Explanations (LIME)

One prominent example that was explored was the LIME model or (Local Interpretable Model-Agnostic Explanations). This model attempts to explain a classifier in an interpretable and faithful manner by creating an interpretable model locally around the prediction [RSG16]. We believed it to be useful for our explainer model because of its ability to be applied to multiple different classification model types, like image and text, meaning it would optimistically be extremely versatile in the case of the CVDD algorithm. In the original LIME paper, the researchers discuss the success the interpreter has had with SVM text classification so, we speculate we would be able to map this type of classifier to a one class classifier for text.

We chose to use the LIME Recurrent Tabular Explainer model because of its ability to work with neural networks and the ability to handle three dimensional data sets. This model is an explainer for keras-style recurrent neural networks. There were some difficulties with having the LIME interpreter be implemented so we ultimatley failed at this attempt. We determined this could be due to the nature of a LIME model working best with local data instances (as opposed to global) and more robust model architectures. With the custom built architecture of CVDD the LIME interpretation model and other models tested, struggled to understand CVDD data and its results. Therefore, with less time to research a different interpretation model, the researchers decided to fore-go this approach in search of something more appropriate for the given model.

3.2 Attention Mechanisms as Interpretations

During the research of interpretation models that utilize attention mechanisms we found ourselves questioning the CVDD paper's claims to have inherent interpretation from an attention mechanism. There was little to no in-depth exploration of this claim in the paper. When researching how to interpret a model using its attention mechanisms we alternatively found research that alleges that this approach is somewhat of a misconstrued/misunderstood idea. While attention mechanisms are a large portion of performance increase, interpretability is only beginning to assess what computed attention weights actually communicate [JW19]. Researchers have even found that correlations between intuitive feature importance measures (including gradient and feature erasure approaches) and learned attention weights is weak for recurrent encoders [SS19].

When the LIME model failed, we decided to observe changes in the CVDD model decisions by changing attention weighting schemes which will help question the real interpretability of said model using attention mechanisms. We plan to evaluate the overall model outputs of anomalous reviews, compare context vectors, and evaluate the text based off of the probability distributions of their weights using the softmax function. The reason we are using the softmax function is because it will evenly distribute all of our weights as a probability from 0 to 1. According to the CVDD model paper, the heaviest weights should equate to the highest probability which should tell us the importance of that word (i.e. larger probability should means the word is more important to the prediction). We want to explore how much the attention weights are really effecting the outputs of anomalous text review data.

4 Experiments

4.1 Data Set

The data set we will be using for this model will be the IMDB Movie Reviews data sets. The reasoning behind using this data set is that it was predetermined by the existing CVDD model designers so, we will be able to easily see if our results match those of the CVDD researchers when we recreate and validate the model. The IMDB movie review data set is a data set of 50,000 movie reviews.

4.2 Recreation Validation

In order to answer our research question we first had to reconstruct the CVDD model. After successfully cloning and running the model on a local machine we were able to compare the context vectors of our local machine validation and the context vector results of the CVDD paper. We can see these results in table 1 and figure 1.

Context Vector 1		Context Vector 3	
great	excellent	plot	characters
good	well	story	storyline
nice	best	narrative	scenes
superb	terrific	subplots	believable
wonderful	better	twists	tale

Table 1: Results from recreating the model for context vectors 1 and 3

c_1		c_3	
great	excellent	plot	characters
good	superb	story	storyline
well	wonderful	scenes	narrative
nice	best	subplots	twists
terrific	beautiful	tale	interesting

Fig. 1 Compared to model results for context vectors 1 and 3

We can see that the context vectors 1 and 3 are are fairly similar with only a few changes. This is most likely due to variable changes in the models learning however we can say this is a successful reproduction of the CVDD model on a local machine.

4.3 Interpretation Comparison

With our new approach we wanted to observe how changing the model attention mechanism heads and weights caused any significant difference in model outcome. In the code of the CVDD model and supporting paper, the researchers assert that having context vectors that are combined with the self-attention weights will inherently make the model highly interpretable. In order to test this we ran the simulation two times, one with the normal replicated parameters as the Ruff paper including 10 attention heads and an attention size of 150. Then a second time with the smallest attention head and size possible (1 for each). We could not go lower than 1 or else a divide by zero error would result. The attention heads parameter is the number of attention heads in the self-attention module and the attention size is the module's dimensionality. So, it was appropriate to reduce the attention mechanism to the smallest size possible to see the results of these parameter differences. It was not possible to remove the attention mechanism entirely for risk of ruining the CVDD model structure which is dependent on this feature.

The final piece of evaluation will be to observe the softmax values of each attention weight per a specific anomalous review. As we have described before, the attention weights are stated to be the

defining factor of a word being important or not for overall classification. Accordingly, an evaluation of the probability values of these weights will be evaluated by using the following softmax equation:

$$\text{Softmax}(x_i) = \frac{\exp(x_i)}{\sum_j \exp(x_j)} \tag{1}$$

Using the softmax as the evaluation metric is for ease of assessment. The raw attention weights have a large distribution of values. The softmax will assist in making the weights an even distribution between 0 and 1 to easily assess. We extracted the raw attention weights and reviews from the CVDD output HTML files. After running the model with these specifications using the IMDB data set we acknowledge the following results.

4.4 Results

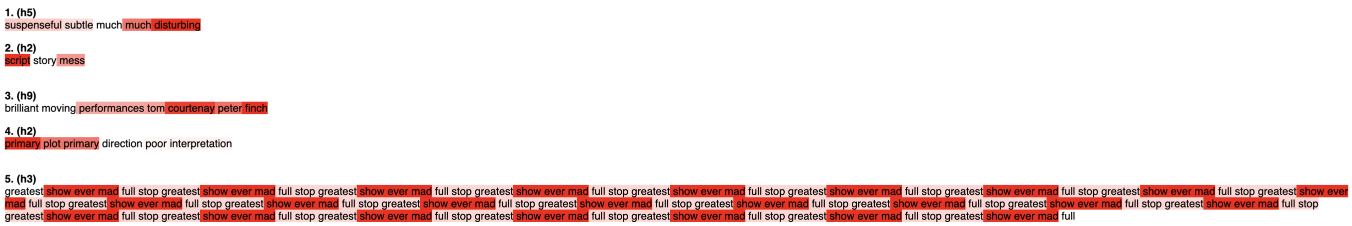

Fig. 3 Results from using CVDD suggested inputs

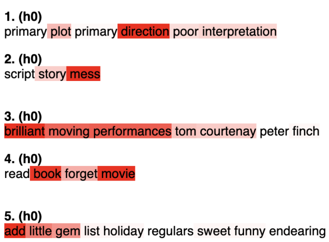

Fig. 4 Results from using value of 1 attention size and heads

We can see the results of the most anomalous reviews are quite different from each other given the attention size differences. We will discuss this more in the next section. Let us also look a few of the top words per context vector to see if there are any changes between them.

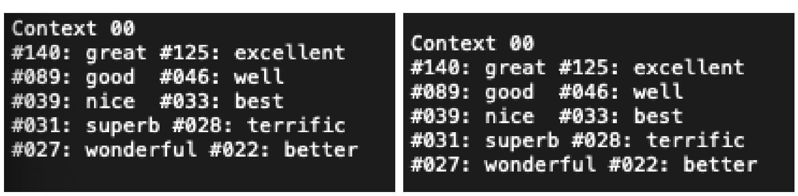

Fig. 5 Context Vector 0 for both calculations

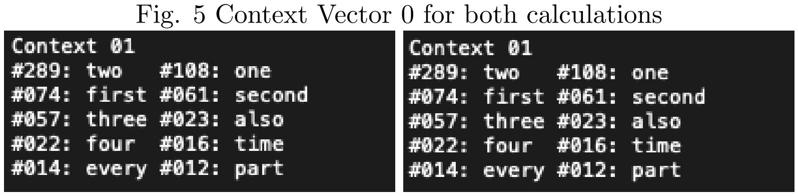

Fig. 6 Context Vector 1 for both calculations

We can clearly see there are no changes between calculations for context vectors 0 and 1 and their top 10 words per context.

For the softmax results we will look at the distribution of review number three for both models "brilliant moving performances tom courtenay peter finch" as these have the same "anomalous" placement but they highlight different sections of the review depending on the attention mechanism's weights. We will show the attention probabilities in a table and plot.

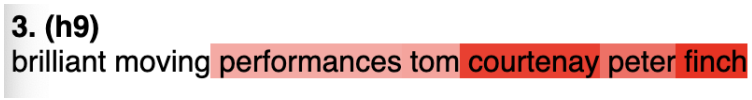

Fig. 7 Large attention mechanism highlighting word attention weight importance

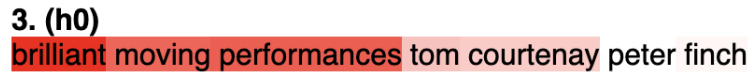

Fig. 8 Small attention mechanism highlighting word attention weight importance

Table 2: Attention Weight Probabilities

Large Mechanism Probabilities	Small Mechanism Probabilities
0.12187871	0.14879098
0.12186832	0.18569201
0.12200583	0.13485167
0.17674124	0.15347273
0.12527732	0.12298144
0.18988678	0.12990998
0.1423418	0.12430119

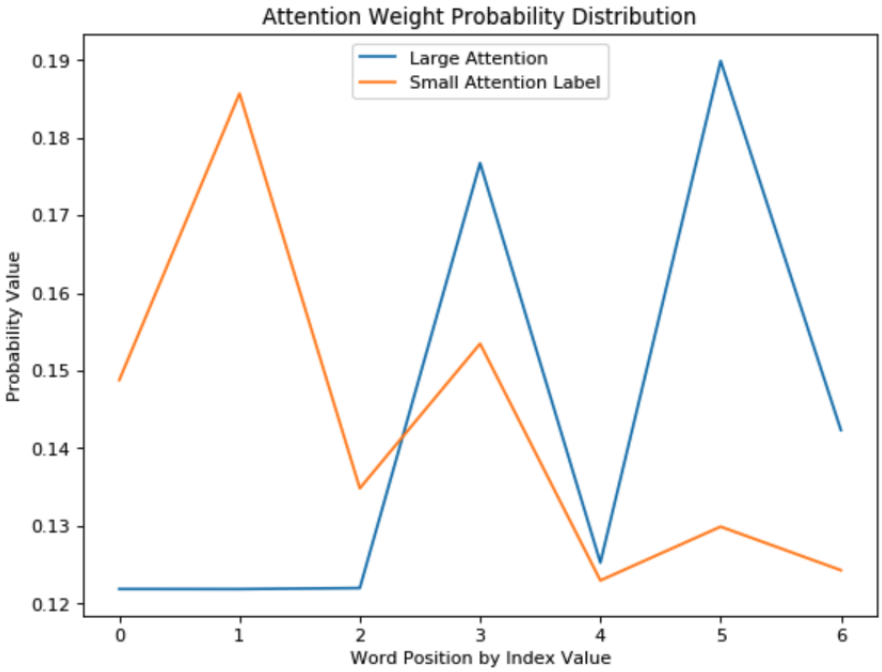

Fig. 9 Plot of Probabilities

5 Discussion

We can see in the results that the top reviews stay the same for the most part but the highlighted words in the reviews in Fig. 3 and Fig. 4 show different words highlighted as the "highest attention weight". The differences in which reviews were chosen is normal given the change in the attention mechanism. As we have just stated, for the reviews that were both chosen for each simulation, there were changes in which words were highlighted as most anomalous. Again this does not necessarily mean anything ground breaking because it could just be due to a change in the model's parameters. That is why we explore more in depth the distribution of these weights in the last part of our discussion.

When looking at the context vectors in figures 5 and 6 we can see that the context vectors stayed the exact same regardless of parameters. This lead us to consider that the context vectors and attention

weights are not as correlated as one previously believed. This indicates to us that the attention mechanism may not even need to be used to find similar contexts. Once again showing that the model's attention weights do not necessarily have a the proposed relationship with the context vectors or have a large effect on the decisions of this model as Ruff et. al had previously implied.

Finally, analyzing the plot of probabilities per word and we can see they are drastically different, just as we were able to see in the important highlighted words of the reviews. We can see there are some inconsistencies with highlighted words vs probabilities in figure 9. For example, with the large attention mechanism we can see that while words 0 and 1 ("brilliant", "moving") have low probabilities, the word "performances" has the same probability value even though it is highlighted as important. For the same mechanism, word positions 3 and 5 have the highest probabilities but they are not the darkest highlighted words. The darkest highlighted words actually have the second and third lowest probabilities even though these words are claimed to be the most important.

We can see a similar pattern with the small attention mechanism probabilities. The darkest highlighted words do not have the highest probabilities. There are inconsistencies with the importance of words and the actual highlighted words. For example, the words "moving" and "performance" have the same shade of highlight yet they do not have the same probability. If these words are deemed the same level of importance, why are they not the same probability level?

The accumulation of these results lead us to believe that the attention weights do not actually have as high of an (or any) impact on word importance to the anomalous text classifier.

5.1 Do Attention Mechanism Really Point to Interpretablity?

When we approached this problem using different interpretation models we were unable to gain any sort of decision results from the CVDD model. Continuing with researching how to find interpretation from attention weights and context vectors for evaluation like the Ruff paper describes, there needed to be additional exploration on the validity of this claim. This led us to wonder if a person can really find interpretation through the attention weights in the model which are supposed to show the importance of each word in the anomalous review. In the Ruff et. al paper it is stated that the highest self-attention weights from the most similar test sentences per-context give us the top words for each context and that the highlighted texts are the most significant attention weights per review [RZV+19]. From the analysis of our results we can see this may not actually be the case. Instead, we observed inconsistencies in the relationships between context vectors outputs and the attention mechanism. As well as, the distribution of weights were not consistent with the proposed importance of each word.

5.2 Conclusion

While we failed to answer our initial research question of being able to build an interpretation model for the Context Vector Data Description model, we did find substantial inference that the claims of interpretablity by design may need some more in depth analysis. Based on analysis of attention weights to the distribution of model outputs, we were able to see that the differences are not that extensive when comparing a large attention mechanism with a small attention mechanism. The link to code used can be found by clicking here.

5.3 Recommendations and Continuing Efforts

This novel approach to interpretation opens a larger space to more deeply analyze the weighting mechanism of this model and other deep learning anomalous text models using an attention mechanism. In order to fully understand the decision choices of CVDD, there must be more time to properly dissect all the aspects of the model. The limitations of this analysis were constrained to deadlines and compiling time that did not allow us to achieve such an in-depth investigation. This is a very surface level exploration of the attention weighting scheme and correlation. However, more comprehensive research can be done to fully understand the relationships in this model. To learn more about the ongoing research of attention weight importance we point to both [SS19] and [JW19] for better in depth research on the relationship between attention mechanisms and the decision making process of a deep learning model.

References

[ASLA20] Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. A diagnostic study of explainability techniques for text classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3256–3274, Online, November 2020. Association for Computational Linguistics.

[ATS20] Christina B. Azodi, Jiliang Tang, and Shin-Han Shiu. Opening the black box: Interpretable machine learning for geneticists. *Trends in Genetics*, 36(6):442–455, 2020.

[CBK09] Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly detection: A survey. 41(3), 2009.

[GBC17] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Introduction*, page 1–1. The MIT Press, 2017.

[HFW08] Kathryn Hempstalk, Eibe Frank, and Ian H. Witten. One-class classification by combining density and class probability estimation. In Walter Daelemans, Bart Goethals, and Katharina Morik, editors, *Machine Learning and Knowledge Discovery in Databases*, pages 505–519, Berlin, Heidelberg, 2008. Springer Berlin Heidelberg.

[IKT] M Ikonomakis, Sotiris Kotsiantis, and V Tampakas. Text classification using machine learning techniques.

[JW19] Sarthak Jain and Byron C. Wallace. Attention is not explanation, 2019.

[KM10] Shehroz S. Khan and Michael G. Madden. A survey of recent trends in one class classification. In Lorcan Coyle and Jill Freyne, editors, *Artificial Intelligence and Cognitive Science*, pages 188–197, Berlin, Heidelberg, 2010. Springer Berlin Heidelberg.

[LRV+20] Philipp Liznerski, Lukas Ruff, Robert A. Vandermeulen, Billy Joe Franks, Marius Kloft, and Klaus-Robert Müller. Explainable deep one-class classification, 2020.

[RSG16] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 1135–1144, New York, NY, USA, 2016. Association for Computing Machinery.

[RZV+19] Lukas Ruff, Yury Zemlyanskiy, Robert Vandermeulen, Thomas Schnake, and Marius Kloft. Self-attentive, multi-context one-class classification for unsupervised anomaly detection on text. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4061–4071, Florence, Italy, July 2019. Association for Computational Linguistics.

[SS19] Sofia Serrano and Noah A. Smith. Is attention interpretable?, 2019.

[SWM17] Wojciech Samek, Thomas Wiegand, and Klaus-Robert Müller. Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models, 2017.

[TS18] M. Thangaraj and M. Sivakami. Text classification techniques: A literature review. *Interdisciplinary Journal of Information, Knowledge, and Management*, 13:117–135, 2018. Copyright - © 2018. This work is published under https://creativecommons.org/licenses/by-nc/4.0/ (the "License"). Notwithstanding the ProQuest Terms and Conditions, you may use this content in accordance with the terms of the License; Last updated - 2022-01-05.

[WRH+18] Qi Wei, Yinhao Ren, Rui Hou, Bibo Shi, Joseph Y. Lo, and Lawrence Carin. Anomaly detection for medical images based on a one-class classification. In Nicholas Petrick and Kensaku Mori, editors, *Medical Imaging 2018: Computer-Aided Diagnosis*, volume 10575, pages 375 – 380. International Society for Optics and Photonics, SPIE, 2018.

[XUD+19] Feiyu Xu, Hans Uszkoreit, Yangzhou Du, Wei Fan, Dongyan Zhao, and Jun Zhu. Explainable ai: A brief survey on history, research areas, approaches and challenges. In *CCF international conference on natural language processing and Chinese computing*, pages 563–574. Springer, 2019.

[ZMH09] Yang Zhang, Nirvana Meratnia, and Paul Havinga. Adaptive and online one-class support vector machine-based outlier detection techniques for wireless sensor networks. In *2009 International Conference on Advanced Information Networking and Applications Workshops*, pages 990–995, 2009.