A Comparison of Unsupervised Abnormality Detection Methods for Interstitial Lung Disease

Matt Daykin[1,2], Mathini Sellathurai[2], and Ian Poole[1]

[1] Canon Medical Research Europe, Edinburgh, UK
[2] School of Engineering and Physical Sciences, Heriot-Watt University, Edinburgh, UK

Abstract. Abnormality detection, also known as outlier detection or novelty detection, seeks to identify data that do not match an expected distribution. In medical imaging, this could be used to find data samples with possible pathology or, more generally, to exclude samples that are normal. This may be done by learning a model of normality, against which new samples are evaluated. In this paper four methods, each representing a different family of techniques, are compared: one-class support vector machine, isolation forest, local outlier factor, and fast-minimum covariance determinant estimator. Each method is evaluated on patches of CT interstitial lung disease where the patches are encoded with one of four embedding methods: principal component analysis, kernel principal component analysis, a flat autoencoder, and a convolutional autoencoder. The data consists of 5500 healthy patches from one patient cohort defining normality, and 2970 patches from a second patient cohort with emphysema, fibrosis, ground glass opacity, and micronodule pathology representing abnormality. From this second cohort 1030 healthy patches are used as an evaluation dataset. Evaluation occurs in both the accuracy (area under the ROC curve) and runtime efficiency. The fast-minimum covariance determinant estimator is demonstrated to have a fair time scaling with dataset dimensionality, while the isolation forest and one-class support vector machine scale well with dimensionality. The one-class support vector machine is the most accurate, closely followed by the isolation forest and fast-minimum covariance determinant estimator. The embeddings from kernel principal component analysis are the most generally useful.

1 Introduction

Distinguishing healthy from diseased anatomy is an important yet challenging task. However, collecting a large amount of labelled ground truth covering diverse pathologies is often impractical or expensive. Unsupervised abnormality detection methods avoid this by enabling the identification of patterns in data that differ from normality. However, anatomy varies substantially between patients and defining a model of normality that reflects these natural variations is not straightforward. To build a normal model, many examples of "normal"

(healthy in our case) instances must be collected and used to learn the distribution of normality. Once the normal model is learnt, novel pathology can be evaluated against this to detect abnormal samples, which, in the medical domain, may be indicative of disease. Collecting examples of normal instances is often an easier task than collecting both normal and abnormal instances, especially where abnormal data may be rare, expensive, or otherwise difficult to obtain.

Several algorithms exist to model normal data in a suitable way. Four of the most influential are the one-class support vector machine (1-SVM) [15], isolation forest (IF) [11], local outlier factor (LOF) [2], and fast-minimum covariance determinant estimator (Fast-MCD) [14].

These four methods come from different families of algorithms: IF is a binary decision tree, 1-SVM is a type of support vector machine, Fast-MCD is a species of Gaussian fit model, and LOF works with local distances. This paper aims to capture the core attributes of these methods and to demonstrate their strengths and weaknesses against one another using CT lung data as a test case. The derived attributes can be used to better understand the family of algorithms each technique represents. We test each method on different embeddings of the data. These embeddings are none (using the raw data), Principal Component Analysis (PCA), Kernel Principal Component Analysis (kPCA), embeddings from a Flat Autoencoder (fAE), and embeddings from a Convolutional Autoencoder (cAE).

The lung data consists of patches of either healthy tissue (normal) or one of four types of pathology (abnormal): emphysema, fibrosis, ground glass opacities, or micronodules.

Our contributions are as follows:

- We compare the performance of four influential abnormality detection methods on interstitial lung data.
- We explore these methods in their accuracy and computational efficiency and discuss the implications of this on their use for medical data.

The paper is organised as follows: prior work in this field (Section 2), a description of the data we use (Section 3), details of the dimensionality-reducing embedding methods we apply to the data (Section 4), an overview of the abnormality detection methods we utilise (Section 5), and finally our results and discussion (Sections 6 and 7).

2 Related Work

Unsupervised abnormality detection appears in several forms. Some algorithms, such as Shared Nearest Neighbour by Ertöz et al. [6] and DBSCAN by Ester et al. [7] seek to cluster data while omitting outliers (defined as data that does not fit into a cluster).

Other work aims to cluster data and then determine the abnormality of each data sample by its distance to the centre of its closest cluster. Kohonen used self-organising maps [10] for this task. Barbará et al. [1] designed a novel algorithm for intrusion detection based on intersecting segments of unlabelled data and

using the intersection as the base data for clustering. Some research has taken a more sophisticated approach where a sample's abnormality score factors in how its local cluster is structured, such as the size of the cluster. He et al. proposed a technique called Find Cluster-Based Local Outlier Factor [8] that uses the size of clusters to influence the decision of abnormality score.

3 Data

We use CT interstitial lung disease data taken from two publicly available datasets - MedGIFT [4] and an emphysema dataset [17].

MedGIFT Dataset: 93 CT scans taken at the University Hospitals of Geneva from patients undergoing high-resolution thorax CT. The scans were acquired with a slice spacing of 10-15 mm and the slices were annotated by a radiologist with 2D regions of interest.

Emphysema Dataset: High-resolution CT scans of a study group of 39 patients (9 never-smokers, 10 smokers, and 20 smokers with chronic obstructive pulmonary disease). Scans were acquired at the Gentofte University Hospital in Denmark and have a within-slice resolution of 0.78 mm. Scans were labelled based on a consensus of an experienced chest radiologist and a CT experienced pulmonologist.

All datasets have segmented ground truth available which is used to define sampling regions and not used beyond this - i.e. the patches used in the experiments are unlabelled. 2D patches of 20 x 20 mm sampled at 1 pixel per mm^2 are extracted from scan slices to capture different pathologies: healthy, emphysema, fibrosis, ground glass opacities, and micronodules in the MedGIFT dataset, and healthy only in the emphysema dataset - see Figures 1 and 2. To maximise the number of healthy samples, the samples taken from the emphysema set are permitted to overlap by 90%. This is a form of data augmentation. The MedGIFT samples have no overlap as there is a sufficient number of patches extracted without augmentation.

The healthy set from the emphysema dataset form a training set, while the MedGIFT samples are divided into two test sets: one containing only the healthy samples, and one containing only the pathological samples. A small random subset is removed from these three datasets for the purpose of parameter optimisation (see Section 5.1). Thus there are three experimental datasets and three optimisation datasets:

Training Set: 5500 healthy ("normal") samples from the emphysema dataset.

Normal Test Set: 1030 healthy ("normal") samples from the MedGIFT dataset.

Abnormal Test Set: 2970 pathological ("abnormal") samples from the MedGIFT dataset: 231 emphysema, 557 fibrosis, 266 ground glass opacities, and 1916 micronodule samples.

Optimisation Training Set: 397 samples removed from the training set.

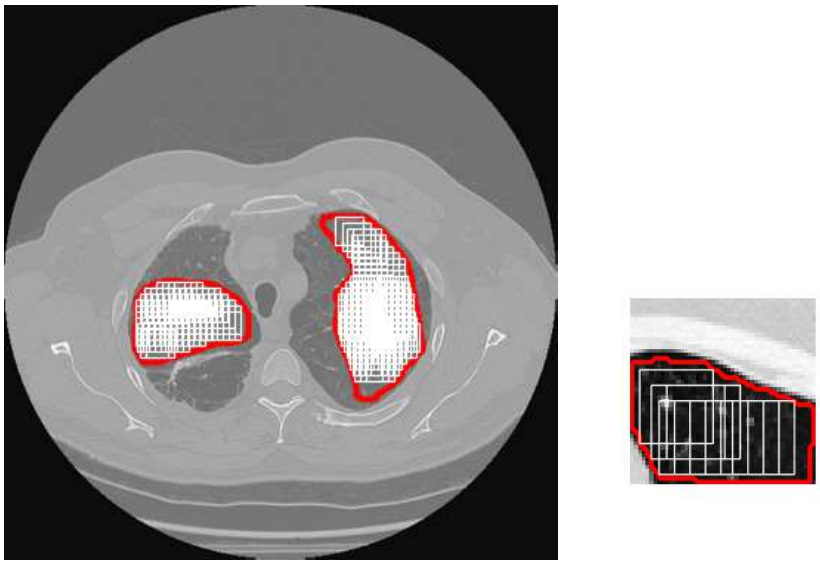

Fig. 1: Left: Example of a lung scan slice showing a sampling region in each lung (outlined region) and extracted patches (boxes). Right: A zoomed example on a small sampling region.

Optimisation Normal Test Set: 203 samples removed from the normal test set.

Optimisation Abnormal Test Set: 585 samples removed from the abnormal test set: 51 emphysema, 114 fibrosis, 42 ground glass opacities, 376 micronodules.

There is no overlap of samples between any of the six datasets. The proportion of each pathology in the abnormal test set derives from the relative abundance of these pathologies in the original patient data.

4 Data Embeddings

To provide a broad understanding of the detection methods, we utilise five embedding methods on the raw data to reduce its dimensionality and capture the salient points of the data. The amount of dimensionality reduction (or number of kept components, x) is tuned to each method - see Table 1.

None : This leaves the data as is with 400 dimensions.

Principal Component Analysis (PCA)[9] : Performs principal component analysis and keeps the x components with the highest variance.

Kernel Principal Component Analysis (kPCA)[16] : Performs kernel principal component analysis and keeps the x components with the highest variance. A third order polynomial kernel is used.

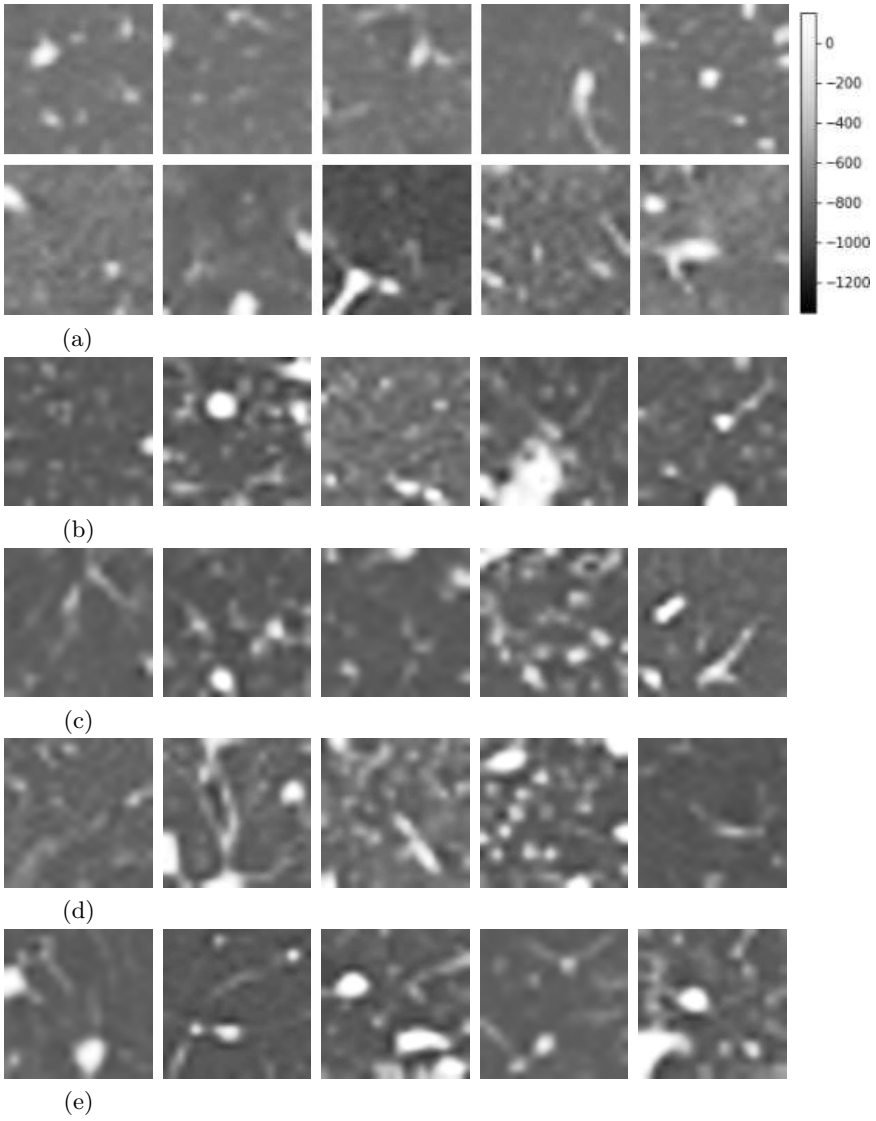

Fig. 2: Examples of extracted patches of lung pathology windowed at a level of -600 HU with a width of 1500 HU, as recommended by Radiopaedia [13]. From top to bottom: a) Healthy (upper row from the training set and lower row from the test set) b) Emphysema c) Fibrosis d) Ground glass opacities e) Micronodules.

Flat Autoencoder (fAE) of six dense layers, the final of which is the decoded output. The encoded data (of dimensionality x) at the central layer are used as input to the abnormality detection methods. Figure 3 illustrates this design. We train for 200 epochs with a batch size of 256, adadelta optimisation [5], and the mean square error loss function.

Convolutional Autoencoder (cAE) of pairs of 2D convolutional and max pooling layers to reduce dimensionality to $25x$ dimensions - see Figure 4. The network architecture constrains the number of dimensions to be a multiple of 25. We train for 200 epochs with a batch size of 256, adadelta optimisation [5], and the mean square error loss function.

The fAE and cAE are implemented through Keras [3] (version 2.1.2).

Table 1: The number of features of samples in the data passed to each abnormality method. These values are obtained by optimising each embedding-method pair on the optimisation data sets, with the exception of the None method, which has a fixed value.

Embedding	**IF**	**Fast-MCD**	**LOF**	**1-SVM**
None (Fixed)	400	400	400	400
PCA	237	347	5	20
kPCA	15	2	3	22
fAE	31	10	40	26
cAE (Multiple of 25)	50	25	25	25

5 Abnormality Detection Methods

Four methods are applied to each dataset embedding:

Local Outlier Factor (LOF)[2] is a nearest-neighbour-based approach that examines the distance to the k^{th} nearest neighbour for each data sample in feature space. This distance is then compared to the distances nearby samples generated to give the final abnormality score. LOF effectively judges each data sample relative to the density of its local area.

One-Class Support Vector Machine (1-SVM)[15] is a support vector machine-based method that transforms the data to a higher dimensional space and seeks to build a hyperplane decision boundary assuming the training points belong to one class and all non-training points belong to another class.

Isolation Forest (IF)[11] is a binary forest approach that at each node randomly selects a dimension and then randomly selects a splitting threshold. It continues until each node has a single sample. An ensemble of trees are constructed using this method. By chance, samples with unusual values are more likely to be isolated early in the tree growing than samples in clusters,

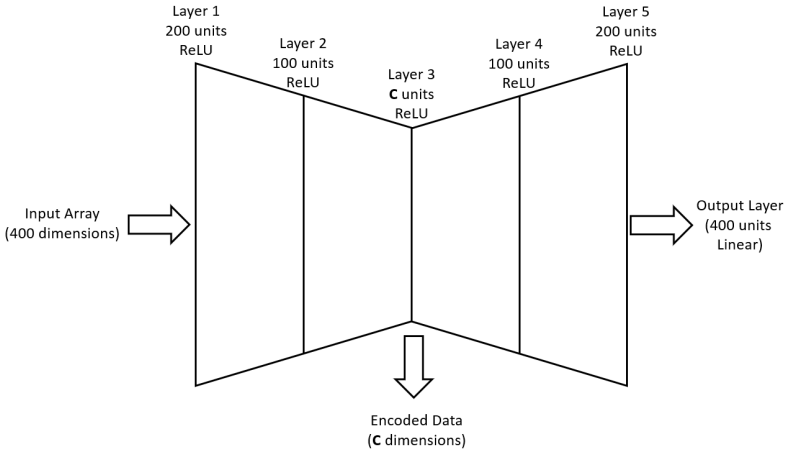

Fig. 3: Structure of the flat autoencoder showing the number of units in each layer and the activation function used. ReLU is the Rectified Linear Unit. Layer 3 has a number of units that varies between the methods (see Table 1).

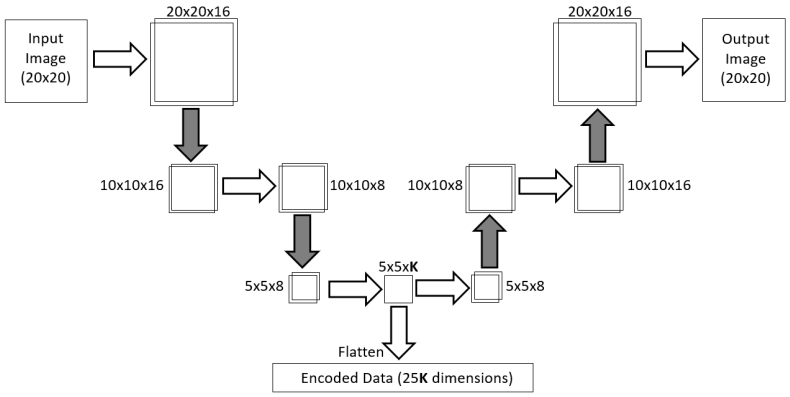

Fig. 4: Structure of the convolutional autoencoder showing the dimensionality after each layer. The horizontal arrows represent 2D convolutions of 3x3 with padding of zero surrounding the convolved image. The shaded arrows represent max pooling or up-sampling (both by a factor of 2) for downwards-facing or upwards-facing arrows respectively. The vertical plain arrow represents flattening of the image for the encoded output as the novelty methods require a 1D input. All layers use the rectified linear unit activation function with the exceptions of the flatten and final layer, which use a linear activation function.

thus the average depth of the sample in trees in the ensemble leads directly to an abnormality score.

Fast-Minimum Covariance Determinant Estimator (Fast-MCD)[14] is a Gaussian fit model that is robust to outliers in the training data, thus leading to a robust fit centred around clusters. The mean and covariance of the nearest fitted Gaussian provide a means of determining the abnormality of any sample in terms of standard deviations from the mean.

These methods are implemented by scikit-learn [12] (version 0.19.1).

5.1 Parameter Optimisation

Each embedding-method pair has a set of parameters to optimise for the problem being solved. We use an optimisation dataset (refer to Section 3 for details) to tune each pair. A gridsearch method is employed to find the optimal parameters in each pair. The following parameters were tuned for each abnormality or embedding method:

- IF: Number of trees in forest. Explored values: 1 - 10 in steps of 1 and 10 - 100 in steps of 10.
- Fast-MCD: Support fraction. Explored values: 0.1 - 0.9 in steps of 0.1.
- LOF: Leaf size and number of neighbours. Explored values: Leaf size: 1 - 10 in steps of 1 and 10 - 100 in steps of 10; Neighbours: 1 - 10 in steps of 1 and 10 - 100 in steps of 10.
- 1-SVM: Nu and gamma. Explored values: Nu: 0.1 - 0.9 in steps of 0.1.; Gamma: 0.1 - 0.9 in steps of 0.1.
- PCA: Number of components. Explored values: first stage: 10 - 390 in steps of 10, followed by second stage: x-9 to x+9 in steps of 1 for the best performing value from the first stage, x.
- kPCA: Number of components and kernel. Explored values: Number of components: first stage: 10 - 390 in steps of 10, followed by second stage: x-9 to x+9 in steps of 1 for the best performing value from the first stage, x; Kernel: *linear*, *poly*, rbf, *sigmoid*, *cosine*. For the *poly* kernel the degree was optimised between 2 and 10.
- fAE: Encoding dimension. Explored values: first stage: 10 - 90 in steps of 10, followed by second stage: x-9 to x+9 in steps of 1 for the best performing value from the first stage, x.
- cAE: Encoding dimension. Explored values: 25-400 in steps of 25.

6 Results

The aim of this evaluation is to demonstrate how the abnormality methods perform on a medical dataset when trained on normal samples taken from one patient cohort and tested on normal and abnormal samples taken from a second cohort. This mimics training a system in a research environment and deploying

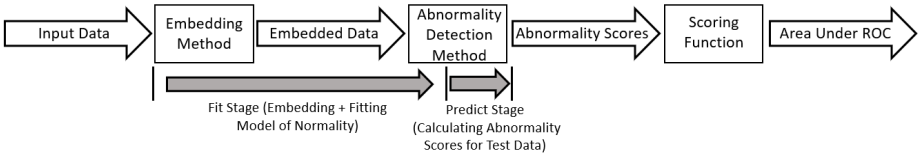

Fig. 5: Our experiment pipeline highlighting the order of data processing and the fit and predict stages. The scoring function takes the abnormality scores and produces an ROC curve from them, from which the area under the ROC curve follows.

it to an external environment, such as a hospital. The objective is to correctly distinguish abnormality from normality and to do this in a time-efficient manner.

Each method is evaluated on the five embedding methods. The methods are trained with healthy patches from one patient cohort and evaluated with a healthy test set and a pathological test set both from a second cohort to produce an abnormality score for all test samples. These scores are used to construct a receiver operating characteristic (ROC) curve. The test scores for the healthy (normal) samples provide the false positive rate, while the test scores for the pathological (abnormal) samples give the true positive rate. The area under the curve is used for the reported accuracy (Table 2). The time required for the algorithm to embed and fit to the data (training stage) and predict abnormality scores (testing stage) is detailed in Table 3[3]. Figure 5 summarises our experiment pipeline.

Table 2: The area under the ROC curve for all embedding methods and abnormality detection methods.

Embedding	**IF**	**Fast-MCD**	**LOF**	**1-SVM**	**Average**
None	0.771	0.657	**0.697**	0.776	0.725
PCA	0.620	0.650	0.620	**0.823**	0.678
kPCA	**0.807**	**0.809**	0.612	0.817	0.761
fAE	0.790	0.771	0.627	0.742	0.733
cAE	0.766	0.758	0.539	0.816	0.712
Average	0.751	0.729	0.619	0.795	

The highest accuracies are for the 1-SVM with a PCA or kPCA embedding where the area under the ROC curve is around 0.82. The IF and Fast-MCD achieve similar to this with accuracies of 0.81 with the kPCA embedding. Overall the kPCA embedding gave the best results. PCA produced the poorest on average. The lowest accuracy is for LOF operating on the cAE embeddings where

[3] Note: The predict times do no include the time taken to run the embedding method on the data being predicted on.

Table 3: The time in seconds required for each experiment combination to fit to the training data and predict on both test sets. Sample size is 5500 patches for fitting and 3000 patches for predicting.

Embedding		**IF**	**Fast-MCD**	**LOF**	**1-SVM**
None	Fit	3.5	69.1	848	18.3
	Pred.	0.9	0.1	678	6.3
	Total	4.4	69.2	1526	24.6
PCA	Fit	2.8	49.0	1.2	2.3
	Pred.	0.6	0.1	0.7	0.7
	Total	3.4	49.1	1.9	3.0
kPCA	Fit	19.1	6.6	5.1	20.4
	Pred.	2.7	2.4	2.7	3.0
	Total	21.8	9.0	7.8	23.4
fAE	Fit	65.1	64.6	71.1	85.5
	Pred.	0.3	0.1	3.6	0.8
	Total	65.4	64.7	74.7	86.3
cAE	Fit	954	1012	1012	959
	Pred.	0.9	0.4	4.6	1.2
	Total	955	1012	1017	960

its accuracy roughly equates to random chance. The LOF method has the lowest accuracy on average and the 1-SVM has the highest.

For the PCA embedding, the IF and Fast-MCD methods are most accurate when using a high number of components (>50%). kPCA used fewer than 6% of the total components for all methods. For the IF and Fast-MCD methods, the kPCA embedding improves the accuracy over PCA. For the LOF and 1-SVM methods there is little difference in the two embeddings.

7 Discussion

Our results demonstrate the key differences between four unsupervised abnormality detection methods when explored in terms of accuracy and speed on CT lung data and embeddings of it. The highest accuracy for each method is:

- IF: 0.807 area under the ROC curve in 21.8 s using kPCA.
- Fast-MCD: 0.809 area under the ROC curve in 9.0 s using kPCA.
- LOF: 0.697 area under the ROC curve in 1525.6 s using the raw data.
- 1-SVM: 0.823 area under the ROC curve in 3.0 s using PCA.

The most accurate method is the 1-SVM, closely followed by the IF and Fast-MCD. LOF has the poorest accuracy. Dimensionality reduction methods have a positive effect on accuracy for Fast-MCD and 1-SVM and a negative impact on LOF. This might be because LOF relies on a distance measure to points in space. If dimensionality is reduced, information on distance between points is lost, and this has a noticeable impact on the accuracy.

These times are the total of the fit and predict times. It is assumed that the fitting stage of the methods would take place in a research environment and the predicting stage in a clinical environment. This means the predict time is much more critical to the successful functioning of the method, providing the fit time remains feasible.

The time taken to complete an experiment is based on two key factors: The speed of the embedding method and the dimensionality of the data.

PCA is the fastest of the tested embedding methods (excluding None), followed by kPCA, then fAE, and finally cAE. For PCA, all experiments are faster than using no embedding, meaning the speedup from the dimensionality reduction outweighed the slowdown from performing PCA for all abnormality methods. kPCA is the most accurate, on average, of all the methods, and PCA is the least accurate. This may be due to non-linear structure present in our data that can only be captured effectively by kPCA. The lower or comparable number of dimensions selected for kPCA relative to PCA supports this idea - kPCA is more efficiently capturing the information in the data.

The LOF method is the most affected by the number of dimensions due to the exponential increase in the number of calculations required with increasing dimensions. The IF method is the least affected as its calculation has no dependency on the number of dimensions.

8 Conclusion

There are a wide variety of algorithms that exist for abnormality detection. This work reviewed four of the most influential algorithms, each representing a different family of solutions: isolation forest, fast-minimum covariance determinant estimator, local outlier factor, and one-class support vector machine. The capability of each of these methods was determined on CT interstitial lung pathology imaging data and five embeddings of it: none (raw data), salient components from principal component analysis, salient components from kernel principal component analysis, embeddings from a flat autoencoder, and embeddings from a convolutional autoencoder. The aim was to correctly distinguish abnormality from normality and to do this in a time-efficient manner.

We showed that local outlier factor had the lowest accuracy and was poorly suited to datasets with a large number of dimensions. The fast-minimum covariance determinant estimator showed better scaling with the number of dimensions but the effect was still noticeable. The isolation forest and one-class support vector machine were the least affected by the number of dimensions. The one-class support vector machine was the most accurate, closely followed by the isolation forest and fast-minimum covariance determinant estimator. Kernel principal component analysis was the most effective embedding technique, leading to the highest average accuracy, but the effect of each embedding varied across the methods.

References

1. Daniel Barbará, Yi Li, Julia Couto, Jia-Ling Lin, and Sushil Jajodia. Bootstrapping a data mining intrusion detection system. In *Proceedings of the 2003 ACM symposium on Applied computing*, pages 421–425. ACM, 2003.
2. Markus M Breunig, Hans-Peter Kriegel, Raymond T Ng, and Jörg Sander. LOF: identifying density-based local outliers. In *ACM sigmod record*, volume 29, pages 93–104. ACM, 2000.
3. François Chollet et al. Keras. https://github.com/keras-team/keras, 2015.
4. Adrien Depeursinge, Alejandro Vargas, Alexandra Platon, Antoine Geissbuhler, Pierre-Alexandre Poletti, and Henning Müller. Building a reference multimedia database for interstitial lung diseases. *Computerized medical imaging and graphics*, 36(3):227–238, 2012.
5. John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(Jul):2121–2159, 2011.
6. Levent Ertöz, Michael Steinbach, and Vipin Kumar. Finding topics in collections of documents: A shared nearest neighbor approach. In *Clustering and Information Retrieval*, pages 83–103. Springer, 2004.
7. Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Kdd*, volume 96, pages 226–231, 1996.
8. Zengyou He, Xiaofei Xu, and Shengchun Deng. Discovering cluster-based local outliers. *Pattern Recognition Letters*, 24(9-10):1641–1650, 2003.
9. Ian T Jolliffe. Principal component analysis and factor analysis. In *Principal component analysis*, pages 115–128. Springer, 1986.
10. Teuvo Kohonen, MR Schroeder, TS Huang, and Self-Organizing Maps. Springer-verlag new york. *Inc., Secaucus, NJ*, 43:2, 2001.
11. Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. Isolation forest. In *Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on*, pages 413–422. IEEE, 2008.
12. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.
13. Radiopaedia. Windowing (ct), 2018. https://radiopaedia.org/articles/windowing-ct.
14. Peter J Rousseeuw and Katrien Van Driessen. A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, 41(3):212–223, 1999.
15. Bernhard Schölkopf, John C Platt, John Shawe-Taylor, Alex J Smola, and Robert C Williamson. Estimating the support of a high-dimensional distribution. *Neural computation*, 13(7):1443–1471, 2001.
16. Bernhard Schölkopf, Alexander Smola, and Klaus-Robert Müller. Kernel principal component analysis. In *International Conference on Artificial Neural Networks*, pages 583–588. Springer, 1997.
17. Lauge Sorensen, Saher B Shaker, and Marleen De Bruijne. Quantitative analysis of pulmonary emphysema using local binary patterns. *IEEE transactions on medical imaging*, 29(2):559–569, 2010.