JOHANNES KEPLER
UNIVERSITY LINZ

Author
Lukas Selch
11941656

Submission
Institute of Computational Perception

Thesis Supervisor
Paul Primus

November 2023

Semi-Supervised Anomaly Detection in Respiratory Sounds: A Comparative Study of Reconstruction and Density Estimation Methods

Bachelor's Thesis

to confer the academic degree of

Bachelor of Science

in the Bachelor's Program

Artificial Intelligence

JOHANNES KEPLER UNIVERSITY LINZ
Altenberger Straße 69
4040 Linz, Austria
jku.at

Abstract

Space for your abstract.

Contents

Chapter 1

Introduction

Respiratory diseases are a leading cause of premature mortality worldwide. With over four million annual deaths attributed to these diseases, early identification and treatment efforts are imperative [10]. The use of chest auscultation, a technique in which respiratory sounds are analyzed with instruments like stethoscopes, is a simple and effective way for diagnosing respiratory diseases.
Automated systems for detecting sound anomalies have become of increasing relevance in the medical field and are driving machine learning research [3]. They have the potential to improve diagnostic accuracy for healthcare professionals and provide initial assessments for patients, ultimately leading to more efficient allocation of healthcare resources.

1.1 Motivation

1.2 Objectives and Approach

1.3 Outline

Chapter 2

Theoretical Background

2.1 Respiratory Sounds

The respiratory system, comprising the airways and lungs, is responsible for the vital function of gas exchange in the human body. Respiratory sounds are generated by the airflow within this system during the inhalation and exhalation [8]. Because these sounds are known to be of great importance in the detection of respiratory pathology, listening to the sounds of breathing through the chest using a stethoscope, called *auscultation*, is a cost-effective, non intrusive, and common part of the physical examination [3].
The characteristics of the sounds observed during chest auscultation can be precisely defined, allowing for a clear distinction between normal and pathological sounds. Normal breathing sounds are heard throughout the inhalation phase, but only at the very beginning of the exhalation phase, and have a relatively narrow frequency band from 100 Hz to 1000 Hz. While there are a variety of different abnormal breathing sounds, we will focus on two of the most prominent and easily recognizable features.
First, *Wheezes* are musical sounds of long duration over 100 milliseconds and of sinusoidal oscillations that can occur during expiration and inspiration caused by airway narrowing or restriction. They range from 100 Hz to 1000 Hz, with higher harmonics possible above that.
The second sound of relevance are *Crackles*. These non-musical sounds are brief and indicate occasional airway opening, possibly caused by secretions. They can be further subdivided into Fine Crackles and Coarse Crackles, which differ in frequency and duration. While Fine Crackles have a characteristic frequency of about 650 Hz and last about 5 milliseconds, Coarse Crackles are longer sounds of more than 15 milliseconds and occur at lower frequencies below 350 Hz [3].
Knowing what constitutes physiological and pathological pulmonary sounds, it is possible to pave the way for automated systems to detect them. Electronic stethoscopes can convert the sound signals from the lung to digital signals, allowing the utilization of advanced anomaly detection algorithms for computer-aided medical diagnosis.

2.1.1 Digital Representation of Sound

Waveform

Mel-Spectrograms

Mel-Frequency Cepstral Coefficients (MFCCs)

2.2 Fundamentals of Anomaly Detection

Anomaly detection addresses the task of identifying deviations in data from anticipated patterns, commonly termed as anomalies or outliers [5]. These anomalies often signal deviations of a system from the norm that are potentially critical and require intervention by the system user. In healthcare, as discussed in section 2.1, outliers in patient respiratory sound patterns can indicate certain conditions.
To find these outliers, anomaly detection algorithms typically define a certain concept of what is considered normal, and any data point that deviates is flagged as an anomaly. The major challenge here is that anomalies are rare and usually not available on a large scale, resulting in a huge data imbalance. In addition, the distribution of anomalies remains uncertain, and even within the set of outliers there may be substantial variation, as data can exhibit various forms of non-normality [19]. In addition, it is important to strike a balance between false positives and false negatives based on the specific application domain, as the severity of one over the other can vary significantly. Outlier detection systems must be properly calibrated to match the characteristics of the domain.
Anomaly detection includes different modes depending on the available data. The simplest, but less common in practical scenarios, is *supervised anomaly detection*. In this approach, a fully labeled dataset is used and the task of the detection system is to determine the boundary between normal and abnormal data points. However, it's important to note that real-world datasets often lack comprehensive coverage of different outlier types, making *unsupervised anomaly detection* more prevalent. In unsupervised anomaly detection, only data points known to be normal are provided, and the system must autonomously learn their defining characteristics. Between these two extremes, there are intermediate modes. In our research, we will address the challenge of *weakly supervised* data, where primarily normal data points are available, but a subset of anomalies is also included to evaluate the real-world effectiveness of the learned representation within a detection system.
Outlier detection algorithms typically output a label that directly classifies the data point as normal or anomalous, or they output an anomaly score, which is a measure of the degree of deviation from normality. This score can then be used to define a threshold above which samples are considered anomalous. In addition to traditional anomaly detection methods such as k-nearest neighbor (KNN) or support vector machines (SVMs), which rely on distance metrics or decision boundaries in the feature space to identify outliers, the application of deep learning methods to anomaly detection has gained traction recently. Traditional methods such as KNN work by identifying the closest data points in a data set and flagging those that are significantly farther away as anomalies. SVMs, on the other hand, focus more on defining a boundary between classes and are particularly effective in scenarios where the separation is clear and well-defined [5]. However, while these methods have proven to be highly effective in many applications, they can be limited in a setting with complex, high-dimensional and noisy data, or data where the anomalies are very

similar to the normal.
Deep learning methods, on the other hand, have the ability to learn and extract features from raw input data and have shown a remarkable ability to effectively learn patterns in complex data structures such as audio signals. In the following, we will delve into two distinct deep learning architecture families that will be employed in this research.

2.3 Reconstruction-Based Methods

In the context of anomaly detection, reconstruction-based methods use reconstruction errors as anomaly scores. The process unfolds in two main steps. First, the method reduces the dimensionality of the original data by transforming it into a latent, more compact representation. Ideally, this *latent space* captures the essential features of the data while dropping noise and unimportant information. Then follows the actual reconstruction where an attempt is made to receive back the original data from the compact representation. The challenge is to find a dimensionally reduced representation that is as compact as possible, while retaining enough essential information from the original data to allow accurate reconstruction and avoid overfitting.
During the training process, the input to a reconstruction-based model contains only data points that are known to be normal, making the process unsupervised. This is critical for the model to learn the typical patterns of standard data. Throughout training, a *reconstruction error* is calculated, which is a measure of how well the reconstructed data matches the original data. In other words, this error evaluates how accurately the model can recreate the input data. The basic assumption is that a model trained exclusively on normal data will be able to reconstruct the original data with minimal error. Consequently, when the trained model encounters anomalous data that deviates from the normal patterns, it will struggle to reconstruct it, resulting in a higher reconstruction error. The magnitude of the reconstruction error can be used as an anomaly score, where a higher error indicates a higher probability that a data point is anomalous and vice versa, allowing the model to be used as an anomaly detection system.

2.3.1 Essentials of Autoencoders

Autoencoders are one way to address the dimensionality reduction and are a popular choice of neural network architecture for reconstruction-based methods, as they are capable of efficiently learn a compact representation of data in an unsupervised fashion. An autoencoder consists of two key components: the Encoder (denoted as $A : \mathbb{R}^n \to \mathbb{R}^l$) and the Decoder (denoted as $B : \mathbb{R}^l \to \mathbb{R}^n$).
The Encoder is responsible for mapping high-dimensional input data of dimensionality n into a lower-dimensional latent space of dimensionality $l < n$. This represents the compression step into a efficient, compact representation of the original data. The Decoder performs the inverse operation where it takes the data from the latent space and reconstructs it back to its original dimensionality, aiming to reproduce the original input as accurately as possible. The result is an optimization problem defined as

$$\arg\min_{A,B} E[\Delta(\mathbf{x}, B \circ A(\mathbf{x}))],$$

where E is the expected value of the reconstruction loss Δ for input $\mathbf{x}$ [2].

2.4 Density Estimation Methods

Probability distributions are mathematical functions that describe the probability associated with each possible value of a random variable. If the random variable can take any value within a certain range, the probability distribution is continuous. It is typically described using a *probability density function* (PDF). The PDF is a fundamental concept in probability and statistics, providing a way to calculate the probability of the random variable falling within a specific interval.
For a continuous real-valued random variable X, the PDF is defined as

$$P(a \leq X \leq b) = \int_a^b f(x)dx$$

for all $a, b \in \mathbb{R}$ [14]. Here, $f(x)$ represents the probability density function of X. It's important to note that the PDF itself does not give probabilities directly. Instead, the probability of X falling within the interval from a to b is given by the area under the curve of the PDF between these two points.
Density estimation is concerned with finding an estimate of the PDF from observed data by estimating a joint distribution $p(\mathbf{x})$ from a set of examples $\{\mathbf{x}^{(t)}\}_{t=1}^{T}$ [12], which helps to understand the distribution of data points within a data set. It can be *parametric* or *nonparametric*. Parametric density estimation assumes that the data follow a certain distribution. The task is then to estimate the parameters of that distribution. For example, in the case of a normal distribution, the mean and variance are to be estimated. Nonparametric density estimation, on the other hand, does not assume a specific distribution for the data and instead attempts to estimate the PDF directly from the data. This allows for more complex distributions, but typically requires more data to produce an accurate estimate.
Density estimation is well suited for anomaly detection tasks. By estimating the PDF of a dataset, it is possible to identify low-probability regions. Data points that fall into these regions can be considered anomalies, making it a straightforward task once the PDF is properly captured.

2.4.1 Introduction to Masked Autoencoders

Masked Autoencoders for Distribution Estimation (MADE) aims to modify autoencoders so that they don't just learn to compress and reconstruct data, but also to understand how the data is distributed. The notion of autoregressive property is essential. While traditional autoencoders do not consider the order of the input data, MADE introduces a mechanism where each part of its output is determined only by the parts of the input that precede it.
To ensure this ordering, MADE masks the weights of each autoencoder layer to control the information flow between neurons of two layers. Each neuron is thereby given a label, a number between 1 and $D-1$, where D is the dimensionality of the input. The masks are then built using this general rule: a neuron in layer l (called k') can only be connected to a neuron in the previous layer $l-1$ (called k) if its label is greater than or equal to the label of k. This can be mathematically expressed as:

$$M^{\mathbf{W}^l}_{k',k} = \begin{cases} 1, & \text{if} m^l(k') \geq m^{l-1}(k) \\ 0, & \text{otherwise.} \end{cases}$$

Here, $M^{\mathbf{W}^l}_{k',k}$ is the weight matrix mask that determines whether a connection is allowed or not. To strictly maintain the autoregressive property, the rule for

the output layer must be slightly modified. Here, a neuron in the last hidden layer can only influence an output neuron if its label is strictly smaller [12].
In summary, MADE provides a solution for controlling the flow of information in an autoencoder so that each part of the output depends only on the preceding parts of the input. The probability of observing the input $\mathbf{x}$ can be calculated as

$$p(\mathbf{x}) = \sum_{d=1}^{D} p(x_d|\mathbf{x}_{<d}).$$

This provides a probabilistic model that can be further used for anomaly detection.

2.5 Evaluation Metrics for Model Comparison

To effectively evaluate and compare the performance of different anomaly detection models, the selection of evaluation metrics is critical. These metrics must not only be relevant to real-world scenarios, but must also provide clear and insightful interpretations that allow for an intuitive understanding of the model's capabilities and effectiveness. In this work, five different metrics are specifically considered, for which the understanding of the concept of the confusion matrix needs to be further explained.
In anomaly detection, a model can attribute a datapoint to be normal or an anomaly. A prediction can have four different outcomes. When the model identifies a data point as normal and it truly is normal, this correct identification is known as a *true negative*. On the other hand, if the model labels a data point as normal but it is actually an anomaly, this incorrect identification is termed a *false negative*. Conversely, when the model predicts a data point as an anomaly and it indeed is an anomaly, we call this a *true positive*. However, if the model mistakenly predicts a normal data point as an anomaly, this error is referred to as a *false positive*. All four outcomes can be summarized in a confusion matrix.

Having introduces this terminology, we focus on defining the used evaluation metrices.

2.5.1 Sensitivity

Sensitivity is the True Positive Rate (TPR), which can be expressed mathematically as

$$\mathrm{TPR} = \frac{\mathrm{TruePositives}}{\mathrm{TruePositives} + \mathrm{FalseNegatives}}.$$

It is the proportion of anomalies correctly identified by the model. It's values range from 0 (not a single anomaly correctly detected) to 1 (perfect anomaly detection), with the goal being to maximize this metric. Especially in medical applications, where missing an anomaly can have serious implications for a patient, this value is desired to be high.

2.5.2 False Positive Rate (FPR)

The FPR is a measure of the proportion of incorrect positive predictions relative to the total number of actual negatives. It is expressed mathematically as

$$\mathrm{FPR} = \frac{\mathrm{FalsePositives}}{\mathrm{TrueNegatives} + \mathrm{FalsePositives}}.$$

It focuses on which normal data points are misclassified as anomalies, ranging from 0 (not a single normal sample is labeled as an anomaly) to 1 (every normal data point is labeled as an anomaly). Lower FPR values are preferred because they indicate a lower rate of false alarms.

2.5.3 Area Under the Curve (AUC)

To evaluate the performance of a model, the trade-off between TPR (benefit) and FPR (cost) is of primary interest and can be plotted on a Receiver Operating Characteristic (ROC) graph. This graph, structured as a unit square, uses the X-axis to plot the FPR and the Y-axis to plot the TPR.

On the unit square, point (0,0) represents a model that predicts all data to be normal, while point (1,1) represents a model that predicts all data to be anomalous. The theoretical goal is to achieve perfect classification, represented by the point (1,0) [9].
The ROC curve is obtained by varying the classification threshold of a model, calculating the TPR and FPR at each threshold, plotting a point on the ROC graph, and then connecting each point with a line. Each point thereby reflects a different trade-off between TPR and FPR.

For a binary classification task such as anomaly detection, a model with random performance represents a diagonal line from (0,0) to (1,1). Any curve above this diagonal is considered better than random, and any curve below is considered worse than random.
To obtain a single value that accurately encodes the insight gained form the ROC, the Area Under the Curve (AUC) is considered. It is the area enclosed by the ROC curve and the x-axis on the unit square and ranges from 0.5 (no discriminative ability of the model) to 1 (perfect classification). It is important to note that the output of models with an AUC below 0.5 can be inverted to obtain better than random classification.
To illustrate the relevance of AUC as a performance measurement, let's consider an example from the field of medicine. In medical diagnostics, a high AUC value is crucial as it indicates the model's ability to accurately distinguish between patients with a specific disease and those who are healthy. This accuracy is vital for ensuring that patients who require treatment are correctly identified and treated, while at the same time avoiding unnecessary medical interventions in healthy individuals. Such a balance is essential in medical practice to provide effective care and minimize harm.

2.5.4 Specificity

Specificity, also known as the True Negative Rate (TNR), is mathematically expressed as

$$\mathrm{TNR} = \frac{\mathrm{TrueNegatives}}{\mathrm{TrueNegatives} + \mathrm{FalsePositives}}.$$

This metric quantifies the proportion of normal data points correctly identified by the model. It's value ranges from 0 (every normal data point is falsely flagged as an anomaly) to 1 (perfect recognition of normal samples), with the goal of achieving a high TNR. In medical diagnostics, maintaining high specificity is essential to avoid false alarms. Together with the Sensitivity, it can be used to calculate the Balanced Accuracy.

2.5.5 Balanced Accuracy (BALACC)

Balanced Accuracy is a performance metric that equally weights the importance of correctly identifying anomalies and normal data points:

$$\mathrm{BALACC} = 0.5 \times (\mathit{Sensitivity} + \mathit{Specificity})$$

TPR and TNR are combined in this metric to provide a more holistic view of model performance, especially in cases where datasets are unbalanced. In a highly unbalanced dataset where one class significantly outnumbers the other, which is often the case in anomaly detection since the anomaly typically occurs less frequently than normal behavior, traditional accuracy can be misleading. For example, in a data set with 95% normal data points and only 5% outliers, a model that always predicts the 'normal' class would achieve 95% accuracy. By averaging both specificity and Sensitivity, the balanced accuracy in this case would be 50%, indicating random performance. This metric has similar goals to AUC, emphasizing the importance of both reliably detecting anomalies and avoiding false alarms. This metric will be used later for better comparison with existing literature.

Chapter 3

Literature Review

3.1 Current State of Respiratory Sound Analysis

Lung auscultation is the standard method of diagnosing respiratory disease by listening to the patient's lungs through the chest. However, this approach, which relies on manual assessment by healthcare professionals, has several limitations. Its effectiveness depends on the physician's skill, experience and auditory sensitivity, leading to potential inaccuracies in diagnosis [18]. In addition, manual auscultation is typically limited to clinical settings, missing critical auditory cues that may occur outside of these settings, such as nocturnal breath sounds common in conditions such as asthma [20].
These limitations, combined with advances in technology, have led to the development of computerized respiratory sound analysis. In this approach, lung sounds are digitally recorded and then analyzed. Early techniques focused on the graphical representation of sound waves, allowing medical professionals to visually identify abnormalities. However, this method did not fully mitigate the risk of human error. Subsequently, statistical approaches were developed to assess the frequency of specific respiratory events based on historical data patterns. According to systematic reviews [18], machine learning based approaches provide the most promising results, but so far were limited by the lack of sufficiently large data sets.

3.1.1 The ICBHI Challenge 2017

During the 2017 Annual International Conference on Biomedical and Health Informatics, a central challenge was launched in response to the scarcity of comprehensive lung sound data. This challenge aimed to foster the development and evaluation of advanced algorithms for automated lung sound classification, using a novel dataset curated specifically for this purpose. Known as the Respiratory Sound Database [21], this collection stands out as one of the earliest and most comprehensive publicly available datasets in the field, comprising 6898 respiratory cycles from 126 patients. These recordings, collected by professional teams in Greece and Portugal, represent a diverse range of audio samples, capturing sounds from healthy individuals as well as patients suffering from lung diseases such as COPD, asthma, and bronchiectasis. Each breathing cycle in the database is annotated by domain experts and categorized as normal, with wheezes, with crackles, or with both wheezes and crackles. The challenge encouraged a multitude of submissions, showcasing a range of innovative machine learning approaches. Below, we compare a selection of these methods.

3.1.2 Existing Approaches

Starting with traditional artificial intelligence methods, Jakovljević and Lončar-Turukalo (2018) published their work based on hidden Markov models (HMM) alongside the paper introducing the Respiratory Sound Database. [16]. Using MFCCs as features, they employed a four-class classifier with the official 60/40 split at the recording level, using 60% of the data for training and the remaining 40% for evaluation. The four classes were healthy, crackles, wheezes, both crackles and wheezes. A balanced accuracy score of 0.39 was achieved, with sensitivity of 0.38 and specificity of 0.41.
Chambres et al. (2018) used boosted decision trees (BDT) to address the four-class classification task [4]. They used the same 60/40 split and MFCCs as features. The model architecture significantly improved the balance accuracy to 0.49, with a sensitivity of 0.78 and a specificity of 0.21.
Shortly after, the use of neural networks gained traction. Ma et al. (2019) proposed the use of a bi-ResNet (LungBRN) architecture [22] consisting of multiple concatenated convolutional neural network layers [17]. Using the same split as the other approaches, but using short-time Fourier transform (STFT) and wavelet analysis to extract features, they achieved an official balanced accuracy of 0.5, specificity of 0.69, and sensitivity of 0.31 for the four-class problem.
The Microsoft Research India team around Gairola et al. (2021) published their RepireNet [11] network and benchmarked it in a variety of data splits and in a binary and four-class classification problem. The backbone are blocks of ResNet34 [15] deep convolutional neural networks (CNN). Using MelSpectrograms as features, their baseline CNN achieved 0.55 balanced accuracy on the official 60/40 split, 0.66 balanced accuracy for the four-class problem on a self-defined random 80/20 split at the breathing cycle level, and 0.72 on the same 80/20 split but treating the problem as a binary classification, which allows for an easier comparison to a anomaly detection setting.

Table 3.1: Performance comparisons of the showcased models

Model	Split	Features	Se	Sp	BALACC
HMM	60/40 4 class	MFCC	0.38	0.41	0.39
BDT	60/40 4 class	MFCC	0.78	0.21	0.49
LungBRN	60/40 4 class	STFT + Wavelet	0.69	0.31	0.5
RespireNet CNN	60/40 4 class	MelSpectrogram	0.39	0.71	0.55
RespireNet CNN	80/20 4 class	MelSpectrogram	0.54	0.79	0.66
RespireNet CNN	80/20 2 class	MelSpectrogram	0.61	0.83	0.72

It is important to note that all mentioned approaches to respiratory sounds analysis rely on treating it as a supervised classification task. These methods, while effective in their context, assume the availability of extensive labeled data representing various specific respiratory conditions. In real-world scenarios, however, such comprehensive data sets are not always readily available. Furthermore, the strict categorization of respiratory sounds into predefined classes may overlook the nuanced and unpredictable nature of respiratory anomalies. Therefore, the remainder of this thesis will explore respiratory sound analysis through the lens of anomaly detection.

3.2 Respiratory Sound Analysis from an Anomaly Detection Perspective

The use of anomaly detection methods to solve sound analysis problems is not a new concept. In particular, these methods have proven their effectiveness in the field of industrial sound analysis, as demonstrated in Task 2 of the annual DCASE Challenge [7], where machine condition monitoring is performed by observing the sound produced by these machines. The sound emitted can be either normal or anomalous, and machine learning algorithms learn to understand the characteristics of healthy machine sounds in order to accurately predict machine failure in the case of anomalous sounds such as rattling or whirring.
A similar approach can be used in breathing sound analysis. The different anomalous respiratory sounds can all be grouped into a single anomaly class, and anomaly detection models can learn the constitution of healthy respiratory cycles. If a sample deviates significantly from the learned representation of a healthy sound, the system can flag it as anomalous.

3.3 Variational Autoencoders

Cozzatti et al. (2022) [6] explored the first anomaly detection approach to the respiratory sound database. In their work, MFCCs were used as features and the breathing cycles containing wheezes, crackles or both were all summarized in an anomaly class. A Variational Autoencoder (VAE) was trained using only known healthy breathing cycles.
Variational Autoencoders are similar to Autoencoders in that they consist of an encoder and a decoder part. The encoder of a VAE, by comparison, uses variational inference to output the parameters of a continuous and easily sampled distribution, usually the mean and standard deviation of a Gaussian [6]. As a result, the input to the decoder is a single sample from that predicted distribution. This allows the model to provide a measure of certainty of the reconstruced data using the variability of the latent space [1].

By training the VAE with normal sounds only, it learns to accurately reconstruct physiological respiratory cycles. The reconstruction error reported by the model is small in this case. When the model attempts to reconstruct a respiratory sound with pathologies, the parameters of the Gaussian will most likely not match the parameters of the learned distribution of healthy sounds, and thus the reconstruction will have a higher error. The paper then used a small subset of the original dataset, containing both healthy and unhealthy lung sounds, to determine a threshold in the reconstruction error above which all higher errors should be marked as anomalous, making the process weakly supervised. The proposed model achieved competitive results in the binary class problem, with a balanced accuracy of 0.57 for the official 60/40 split and 0.6 for a random 80/20 split.

3.4 Group Masked Autoencoders

In section 2.4.1, we have discussed how Masked Autoencoders are an alternative anomaly detection approach to generative models by evaluating probability densities. The basic concept focused on modifying an existing autoencoder

Table 3.2: Performance of the proposed method

Split	Se	Sp	BALACC
60/40	0.33	0.80	0.57
80/20	0.58	0.61	0.60

structure to satisfy the autoregressive property by masking the weights of the neural network layers so that each output dimension depends only on the preceding input dimensions.
When dealing with representations of sound data, such as MelSpectrograms or MFCCS, the interest shifts from the autoregressive ordering of individual input dimensions to the ordering of sound frames. Here, Group Masked Autoencoders (GMADE) [13] provide a more tailored approach for audio anomaly detection tasks where temporal context is important. GMADE differs from traditional MADE in that it does not split the joint distribution into individual dimensional conditions, but rather over grouped frames. This approach is particularly relevant when dealing with sound data, where each time frame in a MelSpectrogram is considered a separate group.
In GMADE, the input space has the dimensionality $T \times M$, where T is the number of frames concatenated in the input and M is the number of Mel frequency bands. If an input sample can be thought of as $\mathbf{t} = [\mathbf{t_{i+1}}, \mathbf{t_{i+2}}, ..., \mathbf{t_{i+T}}]$ with $\mathbf{t_i} \in \mathbb{R}^{M \times 1}$, the joint density can be decomposed as

$$p(\mathbf{t}) = \sum_{i=1}^{T} p(\mathbf{t_i} | \mathbf{t_{<i}})$$

where the probability of each frame depends on all previous frames and their mel bins and no other frames [13]. To maintain the autoregressive property, the generation of the weight matrices must be slightly adapted from the MADE approach to assign labels to the neurons only from 1 to $T - 1$ to correctly zero connections between groups instead of units.
The paper also explored orderings other than causal, where a frame can only depend on its predecessors. Backward ordering predicts the probability density of frames given only their succeeding frames, while middle frame ordering attempts to predict the middle frame given only the frames surrounding it. Ensembles of all three approaches were also evaluated. GMADE achieved state of the art results in Task 2 of the DCASE Challenge 2020 in the machine condition monitoring task, especially for non-stationary sounds. While this is promising for respiratory sound analysis due to the non-stationary nature of lung sounds, the efficacy of GMADE in detecting anomalies in respiratory sounds is yet to be tested.

Chapter 4

Methodology

4.1 Dataset

4.1.1 Definition

4.1.2 Data Splitting

4.2 Preprocessing

4.3 Detailed Overview of the Models

4.3.1 Implementation of the VAE

4.3.2 Implementation of the G-MADE

Chapter 5

Experiments and Results

5.1 Experimental Setup

5.2 Generalization Capabilities

5.3 Age and Gender Differences in Model Accuracy

5.4 Model Sensitivity to Noise

5.5 Model Performance on Crackles and Wheezes

5.6 Impact of Hyperparameter Variations

5.7 Assessment of Data Splitting at Recording Level

5.8 Comparison of Feature Extraction Methods: MFCC vs. MelSpectrogram

5.9 Interpretability and Explainability of the Proposed Models

5.10 Comparative Analysis and Discussion

Chapter 6

Conclusion and Outlook

6.1 Summary of Findings

6.2 Potential Applications and Practical Implications

6.3 Limitations of the Proposed Approaches

6.4 Directions for Future Research

Bibliography

[1] Jinwon An and Sungzoon Cho. 2015. Variational autoencoder based anomaly detection using reconstruction probability. *Special lecture on IE*, 2, 1, 1–18.

[2] Dor Bank, Noam Koenigstein, and Raja Giryes. 2023. Autoencoders. *Machine Learning for Data Science Handbook: Data Mining and Knowledge Discovery Handbook*, 353–374.

[3] Abraham Bohadana, Gabriel Izbicki, and Steve S Kraman. 2014. Fundamentals of lung auscultation. *New England Journal of Medicine*, 370, 8, 744–751.

[4] Gaëtan Chambres, Pierre Hanna, and Myriam Desainte-Catherine. 2018. Automatic detection of patient with respiratory diseases using lung sound analysis. In *2018 International Conference on Content-Based Multimedia Indexing (CBMI)*. IEEE, pp. 1–6.

[5] Varun Chandola, Arindam Banerjee, and Vipin Kumar. 2009. Anomaly detection: A survey. *ACM computing surveys (CSUR)*, 41, 3, 1–58.

[6] Michele Cozzatti, Federico Simonetta, and Stavros Ntalampiras. 2022. Variational autoencoders for anomaly detection in respiratory sounds. In *International Conference on Artificial Neural Networks*. Springer, pp. 333–345.

[7] DCASE. 2023. DCASE2023 Challenge - DCASE — dcase.community. https://dcase.community/challenge2023/index. [Accessed 30-11-2023]. (2023).

[8] J Earis. 1992. Lung sounds. *Thorax*, 47, 9, 671.

[9] Tom Fawcett. 2006. An introduction to ROC analysis. *Pattern recognition letters*, 27, 8, 861–874.

[10] Thomas Ferkol and Dean Schraufnagel. 2014. The global burden of respiratory disease. *Annals of the American Thoracic Society*, 11, 3, 404–406.

[11] Siddhartha Gairola, Francis Tom, Nipun Kwatra, and Mohit Jain. 2021. Respirenet: A deep neural network for accurately detecting abnormal lung sounds in limited data setting. In *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, pp. 527–530.

[12] Mathieu Germain, Karol Gregor, Iain Murray, and Hugo Larochelle. 2015. Made: Masked autoencoder for distribution estimation. In *International conference on machine learning*. PMLR, pp. 881–889.

[13] Ritwik Giri, Fangzhou Cheng, Karim Helwani, Srikanth V. Tenneti, Umut Isik, and Arvindh Krishnaswamy. 2020. Group masked autoencoder based density estimator for audio anomaly detection. In *Detection and Classification of Acoustic Scenes and Events Workshop 2020*. https://www.amazon.science/publications/group-masked-autoencoder-based-density-estimator-for-audio-anomaly-detection.

[14] Charles Miller Grinstead and James Laurie Snell. 2006. *Grinstead and Snell's introduction to probability*. Chance Project.

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778.

[16] Nikša Jakovljević and Tatjana Lončar-Turukalo. 2018. Hidden markov model based respiratory sound classification. In *Precision Medicine Powered by pHealth and Connected Health: ICBHI 2017, Thessaloniki, Greece, 18-21 November 2017*. Springer, pp. 39–43.

[17] Yi Ma, Xinzi Xu, Qing Yu, Yuhang Zhang, Yongfu Li, Jian Zhao, and Guoxing Wang. 2019. Lungbrn: A smart digital stethoscope for detecting respiratory disease using bi-resnet deep learning algorithm. In *2019 IEEE Biomedical Circuits and Systems Conference (BioCAS)*. IEEE, pp. 1–4.

[18] Rajkumar Palaniappan, Kenneth Sundaraj, Nizam Uddin Ahamed, Agilan Arjunan, and Sebastian Sundaraj. 2013. Computer-based respiratory sound analysis: a systematic review. *IETE Technical Review*, 30, 3, 248–256.

[19] Guansong Pang, Chunhua Shen, Longbing Cao, and Anton Van Den Hengel. 2021. Deep learning for anomaly detection: A review. *ACM computing surveys (CSUR)*, 54, 2, 1–38.

[20] Renard Xaviero Adhi Pramono, Stuart Bowyer, and Esther Rodriguez-Villegas. 2017. Automatic adventitious respiratory sound analysis: A systematic review. *PloS one*, 12, 5, e0177926.

[21] BM Rocha, Dimitris Filos, L Mendes, Ioannis Vogiatzis, Eleni Perantoni, Evangelos Kaimakamis, P Natsiavas, Ana Oliveira, C Jácome, A Marques, et al. 2018. A respiratory sound database for the development of automated classification. In *Precision Medicine Powered by pHealth and Connected Health: ICBHI 2017, Thessaloniki, Greece, 18-21 November 2017*. Springer, pp. 33–37.

[22] Runze Wang, Yanan Guo, Wendao Wang, and Yide Ma. 2019. Bi-ResNet: fully automated classification of unregistered contralateral mammograms. In *Artificial Neural Networks and Machine Learning–ICANN 2019: Workshop and Special Sessions: 28th International Conference on Artificial Neural Networks, Munich, Germany, September 17–19, 2019, Proceedings 28*. Springer, pp. 273–283.