

# Deep Learning for Medical Anomaly Detection - A Survey

Tharindu Fernando, Harshala Gammulle, Simon Denman, Sridha Sridharan, and Clinton Fookes

**Abstract**—Machine learning-based medical anomaly detection is an important problem that has been extensively studied. Numerous approaches have been proposed across various medical application domains and we observe several similarities across these distinct applications. Despite this comparability, we observe a lack of structured organisation of these diverse research applications such that their advantages and limitations can be studied. The principal aim of this survey is to provide a thorough theoretical analysis of popular deep learning techniques in medical anomaly detection. In particular, we contribute a coherent and systematic review of state-of-the-art techniques, comparing and contrasting their architectural differences as well as training algorithms. Furthermore, we provide a comprehensive overview of deep model interpretation strategies that can be used to interpret model decisions. In addition, we outline the key limitations of existing deep medical anomaly detection techniques and propose key research directions for further investigation.

**Index Terms**—Deep Learning, Anomaly detection, Machine Learning, Temporal analysis

## I. INTRODUCTION

Identifying data samples that do not fit the overall data distribution is the principle task in anomaly detection. Anomalies can arise due to various reasons such as noise in the data capture process, changes in underlying phenomenon, or due to new or previously unseen conditions in the captured environment. Therefore, anomaly detection is a crucial task in medical signal analysis.

The dawn of deep learning has revolutionised the machine learning field and its success has seeped into the domain of medical anomaly detection, which has resulted in a myriad of research articles leveraging deep machine learning architectures for medical anomaly detection.

The principal aim of this survey is to present a structured and comprehensive review of this existing literature, systematically comparing and contrasting methodologies. Furthermore, we provide an extensive investigation in to deep model interpretation strategies, which is critical when applying ‘black-box’ deep models for medical diagnosis and to understand why a decision is reached. In addition, we summarise the challenges and limitations of existing research, and identify key future research directions, paving the way for the prevalent and effective application of deep learning in medical anomaly detection.

### A. What are Anomalies?

Anomaly detection is the task of *identifying out of distribution examples*. Simply put, it seeks to detect examples that do

not follow the general pattern present in the dataset. This is a crucial task as anomalous observations correlate with types of problem or fault, such as structural defects, system or malware intrusions, production errors, financial frauds or health problems. Despite the straightforward definition, identifying anomalies is a challenging task in machine learning. One of the main challenges arises from the inconsistent behaviour of different anomalies, and the lack of constant definition of what constitutes an anomaly [1], [2], [3]. For example, in a particular context a certain heart rate can be normal, while in a different context it could indicate a health concern. Furthermore, noisy data capture settings and/or dynamic changes in monitoring environments can lead normal examples to appear as out of distribution samples (i.e. abnormal), yielding higher false positive rates [4]. Hence, intelligent learning strategies with high modelling capacity are required to better segregate the anomalous samples from normal data.

### B. Why are Medical Anomalies Different?

The diagram in Fig. 1 illustrates the main stages with respect to medical data processing with machine learning, and how each stage relates to anomaly detection. Collected physiological data is analysed and typically utilised for i) prediction and/or ii) diagnosis. Prediction tasks include predicting future states of physiological signals such as blood pressure, or other characteristics such as recovery rates. For diagnosis tasks a portion of the data is analysed to recognise pathological signs of specific medical conditions. Anomaly detection relates to both prediction and diagnosis tasks, as it captures unique characteristics of the physiological data that could offer information about the data or patient.

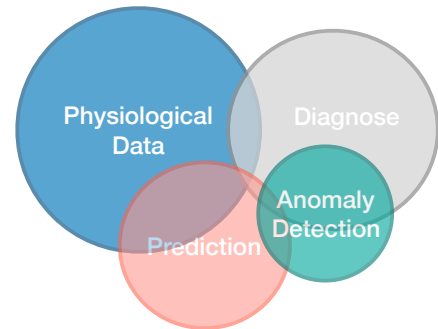


Fig. 1: Illustration of the main stages in medical data processing and how anomaly detection relates to other stages.

Similar to other application domains, medical anomaly detection also inherits the challenges described in Sec. I-A. For

T. Fernando, H. Gammulle, S. Denman, S. Sridharan and C. Fookes are with SAIVT, Queensland University of Technology, Australia (e-mail: t.warnakulasuriya@qut.edu.au.)

instance, Fig. 2 (a) illustrates two examples from the Kvasir endoscopy image dataset [5]. Despite the strong visual similarities, the left figure is an example from the normal-cecum class while the right figure is an example from the ulcerative-colitis disease category. Another example is given in Fig. 2 (b) which illustrates the diverse nature of the normal data in a typical medical dataset. These examples are heart sound recordings from the PhysioNet Computing in Cardiology Challenge 2016 [6]. The top figure shows a clean normal heart sound recording. While the figure in the 2nd row represents a recording of the normal category that has been corrupted by noise during data capture. Therefore, when modelling normal examples, the model should have the capacity to represent the diverse nature of the normal data distribution. Apart from these inherent challenges, medical anomaly detection has additional hindrances which are application specific. Firstly as the end application is primarily medical diagnosis, the test sensitivity (the ability to correctly identify the anomalous samples) is a decisive and crucial factor, and the abnormality detection model is required to be highly accurate. Secondly, there are numerous patient specific characteristics that contribute to dissimilarities among different data samples. For instance, in [7] the authors have identified substantial differences among children from different demographics with respect to their resting state in EEG data. There are also substantial differences between different age groups, genders, etc. Therefore, when designing an accurate medical anomaly detection framework measures should be taken to mitigate such hindrances. Considering these challenges, medical anomaly detection is often posed as a supervised learning task [8], [1], where a supervision signal is presented for the model to learn to discriminate normal from abnormal examples. This is in contrast to other domains such as production defect detection or financial frauds detection, where anomalies are detected in an unsupervised manner.

### C. Why use Deep Learning for Medical Anomaly Detection?

Deep learning is becoming increasingly popular among researchers in biomedical engineering as it offers a way to address the above stated challenges. One prominent characteristics of deep learning is its ability to model non-linearity. Increasing non-linearity in the model can better segregate normal and anomalous samples, and better model the inconsistencies in the data. An additional merit that deep learning brings is its automatic feature learning capability. The availability of big-data [9] and increased computational resources has empowered deep learning's hierarchical feature learning process, avoiding the need to explicitly hand-craft and define what constitutes an anomaly. Another interesting trait of deep learning is its ability to uncover long-term relationships within the data seamlessly through the neural network architecture [1], without explicitly defining them during feature design. For instance, recurrent architectures such as Long Short-Term Memory (LSTM) [10] and Gated Recurrent Units (GRU) [11] can efficiently model temporal relationships in time series data using what is termed 'memory'.

### D. Our Contributions

Although several recent survey articles [3], [2] on anomaly detection have briefly touched upon the medical anomaly detection domain, and despite numerous survey papers published on specific medical application domains [12], [13], [14], [15], [16], there is no systematic review of deep learning based medical anomaly detection techniques which would allow readers to compare and contrast the strengths and weakness of different deep learning techniques, and leverage those findings for different medical application domains. This paper directly addresses this need and contributes a thorough theoretical analysis of popular deep learning model architectures, including convolutional neural networks, recurrent neural networks, generative adversarial networks, auto encoders, and neural memory networks; and their application to medical anomaly detection. Furthermore, we extensively analyse different model training strategies, including unsupervised learning, supervised learning and multi-task learning.

Moreover, this paper provides a comprehensive overview of deep model interpretation strategies that can be used to interpret model decisions. This analysis systematically illustrates how these methods generates model agnostic interpretations, and the limitations of these methods when applied to medical data.

Finally, this review details the limitations of existing deep medical anomaly detection approaches and lists key research directions, inspiring readers to direct their future investigations towards generalisable and interpretable deep medical anomaly detection frameworks, as well as probabilistic and causal approaches which may reveal cause and effect relationships within the data.

### E. Organisation

In Sec. II we illustrate different aspects of deep anomaly detection algorithms, illustrating the motivation for these architectures, and highlighting the complexities associated with medical anomaly detection. Specifically, Sec. II-A illustrates the types of data available in the medical anomaly detection domain, and how different deep learning architectures are designed to capture information from different modalities. Sec. II-B categorises deep anomaly detection architectures based on their training objectives, discussing the theories behind these algorithms and the merits and deficiencies of them. Sec. II-C provides an overview of key application domains to which deep medical anomaly detection has been applied. In Sec. III we theoretically outline deep model interpretation strategies which are a key consideration when deploying deep models in medical applications. Sec. IV illustrates some of the challenges and limitations of existing deep anomaly detection frameworks, and provides future directions to pursue. Sec. V contains concluding remarks.

## II. DETECTING MEDICAL ANOMALIES WITH DEEP LEARNING

In this section we identify different aspects of deep medical anomaly detection algorithms, including the types of data used, different algorithmic architectures, and different application

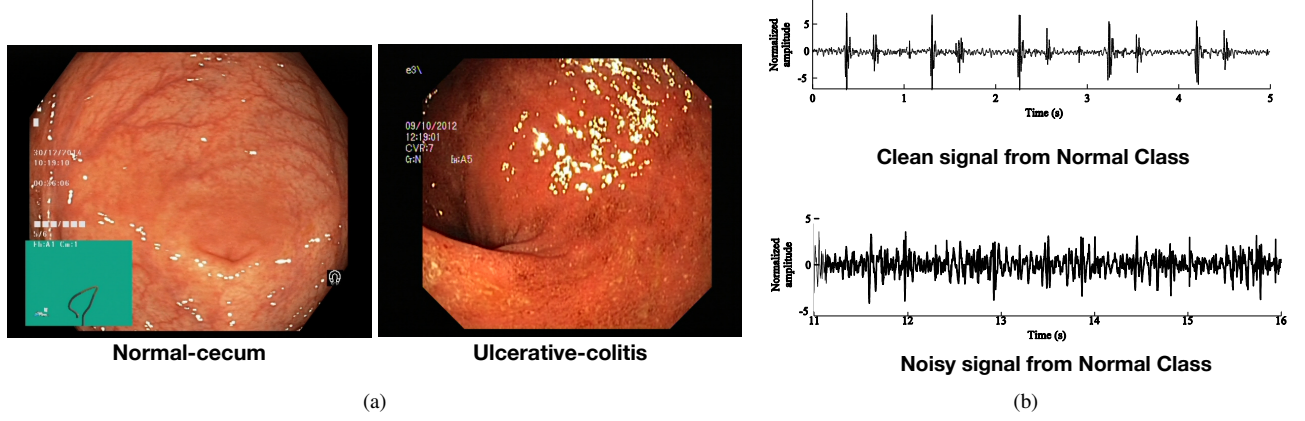


Fig. 2: Challenges associated with medical anomaly detection. (a) Two examples (normal and abnormal) from the Kvasir endoscopy image dataset [5] with strong visual similarities. (b) Two normal examples from the PhysioNet CinC 2016 heart sound dataset [6], where the signal in the bottom row is corrupted by noise.

domains that are considered. The following subsections discuss existing deep medical anomaly detection algorithms within each of these categories.

#### A. Types of Data

Biomedical signals can be broadly categorised into biomedical images, electrical biomedical signals, and other biomedical data such as data from laboratory results, audio recordings and wearable medical devices. The following subsections provide a brief discussion of popular applications scenarios. We also refer the readers to supplementary material where we provide a more comprehensive discussion regarding each of these categories.

##### 1) Biomedical Imaging:

**X-ray radiography:** X-rays have shorter wave lengths than visible light and can pass through most tissue types in the human body. However, the calcium contained in bones is denser and scatters the x-rays. The film that sits on the opposite side of the x-ray source is a negative image such that areas that are exposed to more light appear darker. Therefore, as more x-rays penetrate tissues such as lungs and muscles, these areas are darkened on the film and the bones appear as brighter regions. X-ray imaging is typically used for various diagnostic purposes, including detecting bone fractures, dental problems, pneumonia, and certain types of tumor.

**Computed Tomography scan (CT):** In CT imaging, cross sectional images of the body are generated using a narrow beam of x-rays that are emitted while the patient is quickly rotated. CT imaging collects a number of cross sectional slices which are stacked together to generate a 3 dimensional representation, which is more informative than a conventional X-ray image. CT scans are a popular diagnostic tool when identifying disease or injury within various regions of the body. Applications include detecting tumors or lesions in the abdomen, and localising head injuries, tumors, and clots. They are also used for diagnosing complex bone fractures and bone tumors.

**Magnetic Resonance Imaging (MRI):** As the name implies MRI employs a magnetic field for imaging by forcing protons in the body to align with the applied field.

Specifically, the protons in the human body spin and create a small magnetic field. When a strong magnetic field such as from the MRI machine is introduced, the protons align with that field. Then a radio frequency pulse is introduced which disrupts the alignment. When the radio frequency pulse is turned off the protons discharge energy and try to re-align with the magnetic field. The energy released varies for different tissue types, allowing the MRI scan to segregate different regions. Therefore, MRIs are typically used to image non-bony or soft tissue regions of the human body. Comparison studies have shown that the brain, spinal cord, nerves and muscles are better captured by MRIs than CT scans. Therefore, MRI is the modality of choice for tasks such as brain tumor detection and identifying tissue damage.

In addition to these popular biomedical imaging sensor categories there exists other common data sources such as Positron Emission Tomography (PET), Ultrasound and Medical Optical Imaging. An illustration of different medical imaging signal types is provided in Fig. 3. In the Sec. 1.1 of supplementary material we provide a comprehensive discussion of these different data sources, including a discussion regarding their recent applications in deep learning as well as a list of publicly available datasets.

##### 2) Electrical Biomedical Signals:

**Electrocardiogram (ECG):** ECG is a tool to visualise electricity flowing through the heart which creates the heart beat, and starts at the top of the heart and travels to the bottom. At rest, heart cells are negatively charged compared to the outside environment and when they become depolarized they become positively charged. The difference in polarization is captured by the ECG. There are two types of information which can be extracted by analysing the ECG [17]. First, by measuring time intervals on an ECG one can screen for irregular electrical activities. Second, the strength of the electrical activity provides an indication of the regions of the

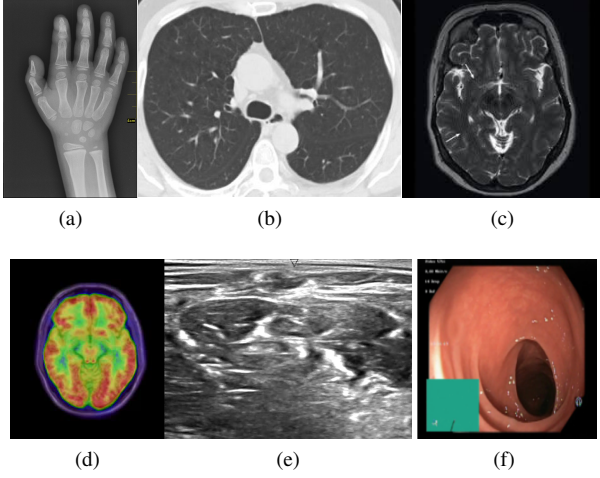


Fig. 3: Illustration of different medical imaging signals. (a) X-ray image (Image Source), (b) Lung CT scans of healthy and diseased subjects taken from the SARS-CoV-2 CT scan dataset, (c) An MRI image with a brain tumor taken from Kaggle Brain MRI Images for Brain Tumor Detection dataset, (d) An example PET scan. Image taken from PET radiomics challenges, (e) An ultrasound image of the neck which is taken from Kaggle Ultrasound Nerve Segmentation dataset, (f) An endoscopy image of the gastrointestinal tract which is taken from The Nerthus endoscopy dataset.

heart that are over worked or stressed.

**Electroencephalogram (EEG):** The EEG detects electrical activity in the brain, which uses electrical impulses to communicate. To capture the electrical activity, small metal discs (electrodes) are placed on the scalp. The electrical signals captured by these electrodes are amplified to better visualise brain activity.

EEGs are a prominent tool for observing the cognitive process of a subject. They are often used to study sleep patterns, psychological disorders, brain damage from head injury, and epilepsy.

In addition to ECGs and EEGs which are the most commonly utilised electrical biomedical signals we would like to acknowledge Electromyography (EMG) sensors where electric potential generated by muscle cells is monitored to diagnose the health of muscles and motor neurons. We refer the readers to Sec. 1.2 of supplementary material for a more comprehensive discussion related to ECGs, EEGs EMGs, and discussion regarding their recent applications in deep learning research and a list of publicly available datasets.

3) *Miscellaneous data types:* In addition to the primary data types discussed above we would like to acknowledge other miscellaneous data sources such as Phonocardiography (PCG) and wearable medical devices which also provide useful information to medical diagnosis. We refer the reader to Sec. 1.3 of the supplementary material where we discuss these data sources in detail, providing discussion related to their recent applications in deep learning research.

## B. Algorithmic Approaches for Medical Anomaly Detection

In this subsection we summarise the existing deep algorithms based on their training objectives, and whether labels for normal/abnormal are provided during algorithm training. In addition, Sec. II-B3 summarises popular recurrent deep neural network architectures used in the medical domain. Finally, a discussion of dimensionality differences between different data types, and how this is managed in existing deep learning research is presented.

1) *Unsupervised Anomaly Detection:* In unsupervised anomaly detection, no supervision signal (regarding whether the example belongs to normal or abnormal category) is provided during training. Therefore unsupervised algorithms do not require labelled datasets and have thus received substantial attention within the machine learning community.

Auto Encoders (AEs) and Generative Adversarial Networks (GANs) are two most common unsupervised deep learning architectures, and are presented in the following subsections.

**Auto Encoders (AEs)** Since their introduction in [18] as a method for pre-training deep neural networks, AEs have been widely used as a method for automatic feature learning [19]. Fig. 4 illustrates the basic structure of an AE. They are symmetric structures and the model is trained to re-construct the input from a learned compressed representation, captured at the center of the architecture.

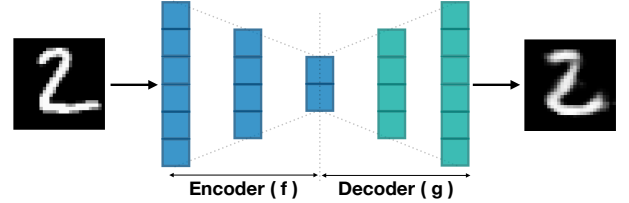


Fig. 4: Illustration of the main components of an Auto Encoder.

Formally, let there be  $N$  samples in the dataset, and the current input be denoted by  $x$ .  $f$  denotes the encoder network while  $g$  denotes the decoder network. Then, the compressed representation,  $z$ , is given by,

$$z = f(x), \quad (1)$$

and reconstructed using,

$$y = g(z). \quad (2)$$

This model is trained to minimise the reconstruction loss,

$$\sum_{x \in N} L(x, g(f(x))), \quad (3)$$

where  $L$  is a distance function, which is commonly the Mean Squared Error (MSE),

$$L_{MSE} = \sum_{x \in N} \|x - g(f(x))\|^2. \quad (4)$$

There exist different variants of AEs. One of such is the *sparse Auto Encoder (S-AE)*. Sparsity constrains the amount



of non-zeros elements in the encoded representation,  $z$ . This can be enforced using an additional penalty term added to the loss (i.e. Eq. 3), which penalises the non-zero elements in  $z$ . One such penalty term is,

$$L_{S-AE} = \sum_{x \in N} L(x, g(f(x))) + \lambda \frac{1}{|N|} \sum_{x \in N} f(x), \quad (5)$$

where  $\lambda$  is a hyper-parameter which controls the strength of the sparsity constraint.

Another category of AEs is *De-Noising AEs* [20], which learn to construct a clean version of the signal from a noisy (corrupted) input. The motivation behind such an architecture is to leverage the de-noising capability to learn a robust and general feature encoding.

*Contractive AEs* are another popular category of AE and try to mitigate the sensitivity of AEs to perturbations of the input samples. A regularisation term is added to the loss defined in Eq. 3 which measures the sensitivity of the learned embedding to small changes of the input. This sensitivity is measured using Frobenius Norm of the Jacobian matrix of the encoder [19].

Finally, the *Variational AE (VAE)* is widely used and is based on the assumption that the observations,  $x$ , are sampled from a probability distribution and it is possible to estimate the parameters of this distribution. Formally, given observations,  $x$ , the VAE tries to approximate the latent distribution,  $P_\phi$ . Let  $\phi$  represent the parameters of the distribution approximating the true latent distribution and  $\theta$  represent the parameters of the sampled distribution, then the objective of the VAE is,

$$L_{VAE}(\theta, \phi; x) = \text{KL}(P_\phi(z|x) || P_\theta(z)) - \mathbb{E}_{P_\phi(z|x)}(\log(P_\theta(x|z))), \quad (6)$$

where KL is the Kullback-Leibler divergence.

There have been a number of applications of auto-encoders for medical anomaly detection. In [21] the authors proposed an AE based method for early detection of respiratory diseases in pigs. The AE model is composed of GRUs to handle the temporal data in the recordings. An EEG based anomaly detection method is proposed in [22] where the authors employ a Convolutional Neural Network (CNN) based architecture for the AE. In contrast, a 3D-CNN is used for the AE in [23] where the authors exploit the 3D feature learning structure of the 3D CNN to handle volumetric CT scans.

In [24] a VAE based framework is proposed to detect anomalies in skin images. In [25] the authors purpose to introduce perturbations to evaluate the effect of input representation variations on the modeled representation. Hence, a two branch structure is proposed where ‘context-dependent’ variations are also added to the VAE branch of the model. This method is validated on an MRI anomaly detection task. Another conditional model is proposed in [26] where the authors condition the VAE output on prior knowledge. The method has been validated on both 2D and 3D anomaly detection tasks.

Despite these interesting characteristics, AEs have limited capabilities when modelling high-dimensional data distributions, often leading to erroneous re-constructions and

inaccurate approximations of the modelled data distribution [27]. Hence, another class of generative models, *Generative Adversarial Networks*, have been introduced.

**Generative Adversarial Networks (GANs):** Another class of AEs are adversarial AEs, more widely known as GANs [28]. They are based on the concept of training two networks, namely a ‘Generator’ (G) and a ‘Discriminator’ (D), which play a min-max game. G tries to fool the D, while D tries avoid being fooled.

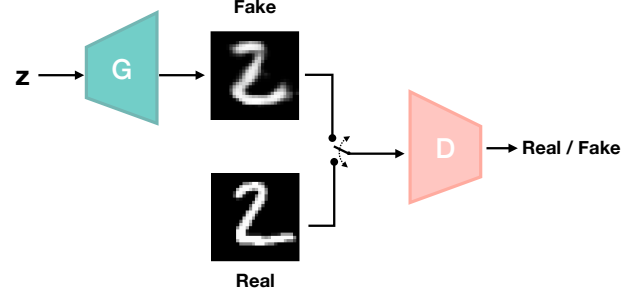


Fig. 5: Main components in Generative Adversarial Networks

Fig. 5 illustrates the basic structure of GAN training. The generator uses noise sampled from  $P_z(z)$  and tries to learn a distribution of the true data,  $P_{data}(x)$ . Specifically, the generator models the mapping from noise space to the data space. The second network,  $D$ , outputs a scalar variable when given a synthesised (fake) or true (real) example. The discriminator is trained to output the correct label of the real/fake classification and this objective can be written as,

$$\min_G \max_D V(D, G) = \mathbb{E}_{x \sim P_{data}(x)}[\log D(x)] + \mathbb{E}_{z \sim P_z(z)}[\log 1 - D(G(z))]. \quad (7)$$

As there is a supervised signal provided to discriminator for training the real/fake classification task, one could assume that GANs are supervised models. However, the real/fake classification is not the primary task and the model is not shown any anomalous examples. Hence, no supervision is given to the GAN framework regarding how to identify abnormalities, making it unsupervised.

A popular sub-class of GANs are *conditional-GANs (cGANs)*, in which both generator and the discriminator outputs are conditioned on additional information,  $c$ . The objective of the cGAN can be written as,

$$\min_G \max_D V(D, G) = \mathbb{E}_{x \sim P_{data}(x)}[\log D(x|c)] + \mathbb{E}_{z \sim P_z(z)}[\log 1 - D(G(z|c))]. \quad (8)$$

cGANs are a popular choice when the synthesised output should depend on a certain environmental context [29], [30].

*Cycle Consistency GANs (Cycle-GANs)* are also a popular variant of GANs, especially for image-to-image translation tasks. Cycle-GANs provide an additional constraint to the framework: that the original input can be synthesised from the generated output.

An example application of GANs for medical anomaly detection is in Schlegl et. al [31] where the authors used

a GAN based framework for anomaly detection in Optical Coherence Tomography (OCT). They first trained a GAN to generate normal OCT scans using the latent distribution  $z$ . Then an encoder is used to map normal OCT scans to  $z$ . Hence, it should be possible to recover an identical image when mapping from the image to  $z$  using the encoder, and from  $z$  to image using the generator. When there are anomalies the authors show that there exist discrepancies in this translation, and propose to identify anomalies using this process.

Despite the appeal of unsupervised approaches, the most common learning paradigm for medical abnormality detection has been supervised learning. We argue that the superior results that supervised training can offer has led to more approaches being developed in this manner.

2) *Supervised Anomaly Detection*: Considering the requirement for a high degree of sensitivity and robustness, particularly due to the diagnostic applications, supervised learning has been widely applied for medical anomaly detection. In contrast to the unsupervised learning structure illustrated in Sec. II-B1, in supervised anomaly detection a supervised signal is provided indicating which examples are from the normal category and which are anomalous. Hence, this is actually a binary classification task and models are typically trained using binary cross entropy loss [32]. This can be formally written as,

$$L = -y \log(f(x)) - (1 - y) \log(1 - f(x)), \quad (9)$$

where  $y$  is the ground truth label,  $f$  is the classifier and  $x$  is the input to the model. Example architectures include the CNN structures in [33] and [34] where they have employed supervised learning to identify anomalies in retina images and for automated classification of skin lesions, respectively. In [35] a deep belief network is trained to detect seizures in EEG data.

*Multi-task Learning (MtL)* is a sub category of supervised learning. The advantage of MtL is sharing relevant information among several related task, rather than learning them individually [36]. For instance, to overcome the challenges which arise due to subject specific variations, a secondary subject identification task can be coupled with the primary abnormality detection task. Hence, the model identifies the similarities and differences among subjects while learning to classify them. Several studies have leveraged MtL in the medical domain. For instance, in [36] the authors purposed an efficient kernel learning structure for multiple tasks and applied this framework to regress Parkinson's disease symptom scores. In [37] a multi-task learning strategy is formulated through detection and localisation of lesions in medical images. The system jointly learns to detect suspicious images as well as to semantically segment the regions of interest in those images.

The deep learning architectures that we discussed so far are feed-forward architecture, i.e. the data travels in one direction only, from input to the output. Despite the simple structure, the deficiency of the feed-forward architecture is its inability to model temporal signals such that the output at a particular time-step can be given as a feedback signal in the next time

step. To address this limitation, Recurrent Neural Networks are introduced.

3) *Recurrent Neural Networks (RNNs)*: Recurrence is a critical characteristic in tasks such as time-series modelling. Recurrent means the output of the current time step is also fed as an input to the next time step. In medical data processing, this is important when modelling sequential data such as EEG and Phonocardiographic data, where we are required to model the temporal evolution of the signal.

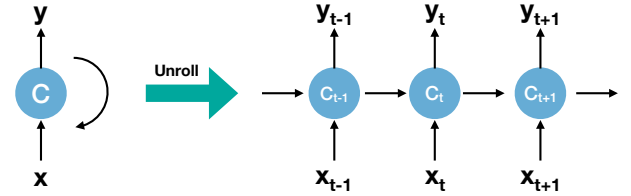


Fig. 6: Illustration of Recurrent Neural Network structure and how it temporally unrolls.

Fig. 6 illustrates the basic structure of an RNN, where we show how it temporally unrolls. Due to this temporal structure, RNNs require Backpropagation Through Time (BPTT) [38], as the gradient of the error of a particular time step depends upon the prediction at the previous time step, and errors accumulate over time. Despite its interesting characteristics, vanishing gradients [39], [40] are one of the major drawbacks of simple RNN structures. This is due to BPTT, and makes them ineffective when modelling long-term dependencies (relationships between distant time-steps).

Several variants of RNN have been introduced to address this limitations. In the following sections we illustrate three of such popular variants, namely, Long Short-Term Memory (LSTM) Networks, Gated Recurrent Units (GRUs) and Neural Memory Networks (NMNs).

*Long Short-Term Memory (LSTM)* [10] networks are specifically designed to model long-term dependencies which are ill represented by RNNs due to vanishing gradients. It introduces a ‘memory cell’ (or cell state) to model long-term dependencies, and a series of gated operations to manipulate the stored information in the memory and update it over time. The core idea behind LSTMs is that the long-term dependencies can be effectively stored in the cell state without losing them when the modelled sequence becomes too long [10]. There are three gates which control what is retrieved from the cell state and what is written back to it.

First, we have ‘forget gate’ which determines what portion of information from the previous time step to retain at the current time step. This gate is controlled by the output at the previous time step and the current input, and the gate has a value between 0 and 1 to control information flow (See Fig. 7). This can be written as,

$$f_t = \sigma(w^f[h_{t-1}, x_t] + b^f), \quad (10)$$

where  $w^f$  and  $b^f$  are the gate’s weights and bias,  $h_{t-1}$  is the previous time step’s output,  $x_t$  is the current input and  $\sigma$  is a sigmoid function.

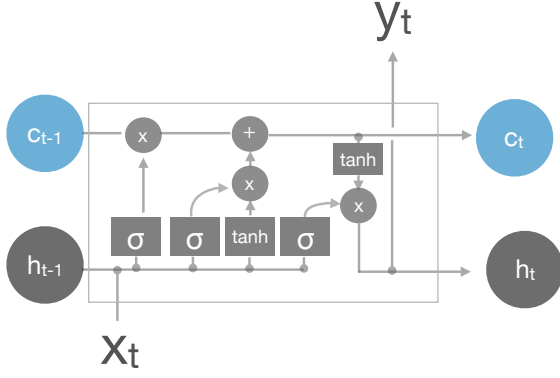


Fig. 7: Visual illustration of Long Short-Term Memory cell.

The next step is to decide what information is written to the cell, which is done via the ‘input gate’. Similar to the previous gate we have a function deciding what portion of information to write,

$$g_t = \sigma(w^i[h_{t-1}, x_t] + b^i), \quad (11)$$

and we use a tanh function to generate the information to write,

$$\tilde{c}_t = \tanh(w^c[h_{t-1}, x_t] + b^c). \quad (12)$$

Then the cell state can be updated using,

$$c_t = f_t \times c_{t-1} + i_t \times \tilde{c}_t. \quad (13)$$

Note how the information from the previous cell state and the current time step are controlled through the forget and input gate values. The final step is to decide what information to output from the cell at the current time step. This is done through the output gate,

$$o_t = \sigma(w^o[h_{t-1}, x_t] + b^o), \quad (14)$$

and,  $h_t$ , the current time step’s output is given by,

$$h_t = o_t \times \tanh(c_t). \quad (15)$$

*Gated Recurrent Units (GRUs)* are another popular variant of the LSTM which were introduced by Cho et. al in 2014 [41]. They combine the forget gate and input gate into a single ‘update gate’. Specifically, Eq. 13 becomes as,

$$c_t = f_t \times c_{t-1} + (1 - f_t) \times \tilde{c}_t. \quad (16)$$

*Neural Memory Networks (NMNs)* are another variant of RNNs, where an external memory stack is used instead of an internal cell state to store information. A limitation of LSTMs and GRUs is that content is erased when a new sequence is presented [42], [43]. This is because such architectures are designed to map temporal relationships within a sequence, not between sequences [42], [43], [44], [45]. Hence, the limited capacity of the internal cell state is not sufficient to model relationships across a large data corpus [46], [42].

Fig. 8 illustrates a typical NMN architecture, which is composed of a memory stack to store the information, and a set of controllers (read, output and write) to provide the memory functionality. We would like to highlight the similarities between the LSTM gated operations and the controller

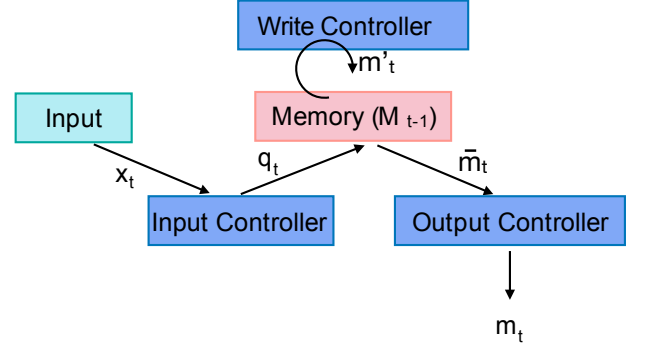


Fig. 8: Illustration of the main components in Neural Memory Network architecture.

functionalities to manipulate the external memory. Specifically, let the state of the external memory,  $M \in \mathbb{R}^{k \times l}$ , at times instance  $t - 1$  be given by  $M_{t-1}$ , where  $l$  is the number of memory slots and  $k$  is the size of each slot. The current input is denoted by  $x_t$ . Then the read controller,  $f^r$ , generates a vector,  $q_t$ , to query the memory,

$$q_t = f^r(x_t). \quad (17)$$

Then using a softmax function we measure the similarity between the content of each memory slot and the query vector such that,

$$z_t = \text{softmax}(q_t^\top M_{t-1}). \quad (18)$$

The resultant score vector,  $z_t$ , denotes the relevance of memory content to the current input. Here, we would like to draw parallels to input gate functionality of an LSTM. The input gate determines what information to extract from the history. However, in contrast to the LSTM architecture, where there is only one single vector storing information related to history, in NMNs there are multiple memory blocks to consider. Hence, attention is employed to extract most relevant information.

Then the output controller,  $f_o$ , generates the output such that,

$$\bar{m}_t = z_t[M_{t-1}]^\top, \quad (19)$$

and,

$$m_t = f^o(\bar{m}_t). \quad (20)$$

This aligns with the output gate functionality in the LSTM. Utilising the input at current time-step and the retrieved historical information from the memory, an output is composed. As the final step the write controller,  $f^w$ , generates a vector to update the memory,

$$m'_t = f^w(m_t), \quad (21)$$

and updates the memory using,

$$M_t = M_{t-1}(I - z_t \otimes e_k)^\top + (m'_t \otimes e_l)(z_t \otimes e_k)^\top, \quad (22)$$

where  $I$  is a matrix of ones,  $e_l \in \mathbb{R}^l$  and  $e_k \in \mathbb{R}^k$  are vectors of ones and  $\otimes$  denotes the outer product which duplicates its left vector  $l$  or  $k$  times to form a matrix [47], [1]. As NMNs are a relatively new concept we refer interested readers to [46].

While the exact memory update mechanisms for LSTMs and NMNs are dissimilar, we would like to highlight the parallels between the LSTM forget gate and the write controller. The LSTM forget gate considers the current time-step's input and the previous cell state (i.e memory) and determines what to pass through from the history. Similarly, the write controller, utilising the NMN output and the retrieved historical information, determines what memory slots to update.

There are numerous works that have utilised RNNs for medical anomaly detection. For instance, RNNs have been utilised in [48] and [49] for text based abnormality detection in electronic health records; and in [50] to detect abnormal heart beats in Phonocardiographic recordings.

More recently, NMNs have been applied in medical anomaly detection, where works have illustrated the utility of external memory storage to memorise the similarities and differences between normal and anomalous examples. Specifically, in [1] the authors couple an NMN together with neural plasticity framework to effectively identify tumors in MRI scans and abnormalities in EEGs. Furthermore, in [51] the same architecture is used to identify different seizure types in EEGs.

### C. Applications

This subsection provides a detailed discussion of popular application domains within deep medical anomaly detection, illustrating how previously discussed architectural variants are leveraged in these domains.

1) *MRI based Anomaly Detection*: In this section we summarize some of key recent literature in deep learning based anomaly detection with MRI data. Fusion of modalities has been a popular research direction which has recently emerged for MRI analysis. In [52] the authors investigate the fusion of T1-weighted (T1w) MRIs and myelin water imaging of MRIs (which is a quantitative MRI technique that specifically measures myelin content) for diagnosis of Multiple sclerosis (MS). In their proposed architecture, they utilise two modality specific Deep Belief Networks (DBN) [53] to extract features from the individual T1w and myelin maps. This is followed by a multi-modal DBN which jointly learns complementary information from both modes. They retrieve a multi-modal feature vector by concatenating the top-level hidden unit activations of the multimodal DBN. As the final step a Random Forest [54] is trained to detect abnormalities. The proposed algorithm is validated using an in-house dataset, which consists of 55 relapse-remitting MS patients and 44 healthy controls. The classification accuracy was  $70.1 \pm 13.6\%$  and  $83.8 \pm 11.0\%$  for T1w and myelin map modalities, respectively, while the fused representation achieved  $87.9 \pm 8.4\%$ .

A strategy to fuse MRI images with fluorodeoxyglucose positron emission tomography (FDG-PET) samples has been proposed in [55]. In their approach, the authors first segment the MRI images into gray and white matter regions, followed by subdivision of the gray matter into 87 anatomical Regions Of Interest (ROI). Then they extract patches of sizes 1488, 705 and 343 from these regions. Similar size patches are also extracted from FDG-PET images. Then 6 independent

Deep Neural Networks (DNNs) with dense layers are used to embed patch information and another DNN is used to fuse the encoded embedding. A softmax layer is used generate the final abnormality classification. The authors utilise this architecture to detect pathologies related to Alzheimer's Disease and the framework is evaluated using publicly available Alzheimer's Disease Neuroimaging Initiative (ADNI) database [56], which contains 1242 subjects. The proposed method achieves 82.93 % accuracy and an approximately 1.5% improvement over utilising FDG-PET alone.

A method to fuse Apparent Diffusion Coefficients (ADCs) of MRIs together with T2-weighted MRI images (T2w) is proposed in [57]. In contrast to predicting a single score level classification, they proposed a method which outputs a segmentation map for each modality, indicating the likelihood of each pixel belonging to the class of interest. They propose to utilise a novel similarity loss function such that the ADC and T2WI streams produce consistent predictions, allowing complementary information to be shared between the streams. The initial segmentation maps are combined with hand-crafted features and passed through an SVM to generate the final predictions. Evaluations were carried out using a dataset with 364 subjects, with a total of 463 prostate cancer lesions and 450 identified noncancerous image patches in which the framework achieves a sensitivity of 89.85% and a specificity of 95.83% for distinguishing cancerous from non-cancerous tissues.

In contrast to the above approach which employs feature level fusion, an architecture using decision level fusion is proposed in [58]. The proposed approach has an ensemble of classifiers composed of 3 convolutional neural networks which are trained separately. Each network provides a softmax classification denoting the likelihood of four Alzheimer's disease classes: non-demented, very mild, mild and moderate. The fusion of the individual classifications is performed using majority voting. The evaluation is conducted on the public OASIS dataset [59] which consists of 416 subjects, and the proposed ensemble method achieves 94 % precision and 93 % recall.

Despite the architectural differences, the above discussed methods are all supervised Deep CNN (DCNN) models and these dominate the MRI based anomaly detection literature. This is clearly motivated by the fact that supervised CNN models are highly effective when extracting task specific spatial information from two dimensional inputs.

Despite the prevalence of supervised DCNN models, a number of approaches have also used Auto-Encoders (AE) [60], [61]. In [61] an AE network with a sparsity constraint has been proposed for the diagnosis of individuals with schizophrenia. First the AE is trained in an unsupervised manner for feature extraction and in the second stage the authors use validation set of the dataset to fine tune the network after adding a softmax layer to the AE. As the final stage a linear support vector machine is used to classify samples. The system is validated using a large scale MRI dataset which is collected from 7 image sources and consists of 474 patients with schizophrenia and 607 healthy controls. This model achieves approximately 85 % accuracy in a k-fold cross validation setting. Similarly,



in [60] an AE is trained for early detection of acute renal transplant rejection. As the first stage the AE is trained in an unsupervised manner. To classify inputs, the decoder of the AE is removed and a softmax layer is trained using supervised learning. This method achieves 97% classification accuracy in a leave-one-subject-out experimental setting on 100 subjects.

Critically, unlike the unsupervised AE models discussed in Sec. II-B1 these models are not purely unsupervised architectures. Rather after the preliminary training of the AE, a classification layer is added and trained using a supervised signal to detect anomalies.

In a different line of work, a multi-scale multi-task learning framework is proposed in [62] for diagnosis of Lumbar Neural Foramina Stenosis (LNFS). Fig. 9 illustrates the architecture used. The authors show that each lumbar spine image can have multiple organs captured at various scales. Furthermore, they illustrate that multi-task learning can be used such that learning from multiple related tasks can boost learning, as discussed in Sec. II-B2, and we note that this strategy is seen across multiple applications domains. Feature maps are extracted at multiple scales and at each level the model tries to perform two tasks, regression of bounding boxes to locate the organs and prediction of abnormalities in the located organs. This system is validated using 200 clinical patients and it is capable of diagnosing abnormal neural foramina with 0.83 precision and 0.8 recall.

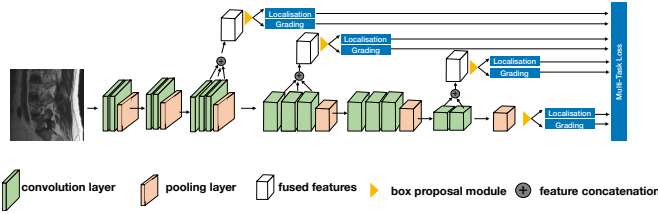


Fig. 9: The architecture proposed in [62] for diagnosis of Lumbar Neural Foramina Stenosis. Recreated from [62]

In addition to those approaches, a Neural Memory Network (NMN) based approach is proposed in [1]. This method utilises the recurrent structure of NMN to compare and contrast characteristics of the samples in the entire dataset using supervised learning. The memory stack stores important characteristics that separates normal and anomalous samples. Therefore, this architecture significantly deviates from rest of the approaches already described. Specifically, a ResNet-50 CNN is used to extract a  $14 \times 14 \times 256$  feature from the input MRI image. This feature becomes the input to the read controller of the NMN. Utilising this as a query vector, the read controller attends to the content that is stored in the memory, comparing them to find the best possible match to the input. The output of the memory read function and the input vector are passed through an output controller which generates the memory output (i.e. a feature vector which is subsequently used to generate the final classification). As the final step, the write controller decides how to update the content in the memory slots, reflecting the information retrieved from the current input. In addition to this typical functionality of the NMN, plasticity is utilised in the NMN controllers such that they can adapt the connectivity

dynamically, changing the overall behaviour of the NMN. This framework was evaluated using the dataset of [63] which contains MRI images captured from 233 patients with different types of brain tumours: meningioma (708 samples), glioma (1426 samples), and pituitary (930 samples). In the 5-fold cross validation setting the model achieves 97.52% classification accuracy. Here we would like to point out that instead of binary normal/abnormal classification, a multi-class classification was conducted where the model discriminates between different abnormal classes using the categorical cross-entropy loss.

2) *Detecting abnormalities in Endoscopy Data:* In this section we summarise some popular deep learning architectures introduced for abnormality detection from endoscopy's.

Considering the fact that endoscopy devices capture RGB data, CNNs pre-trained on large scale object detection benchmarks such as Image-Net [64] have been extensively applied. For instance, in [65] the authors apply the Xception [66] CNN architecture pre-trained on [64] for the detection of ulcers in endoscopy images. The proposed system is evaluated using a dataset that consists of 49 subjects and the authors have performed both 5-fold cross validation and a leave-one-subject out evaluation. The system achieves an average of 96.05 % accuracy in the 5-fold cross validation setting, while the performance varies between 73.7% to 98.2% in the leave-one-subject out evaluation. Similarly in [67] both GoogLeNet [68] and AlexNet [69] pre-trained networks have been investigated for the classification of ulcers. The models were tested on a public dataset [67] which consists of 1875 images. Furthermore, in [70] AlexNet [69] has been applied for both ulcer and erosion detection. The resultant model is capable of achieving 95.16% and 95.34% accuracy levels for ulcer and erosion detection when tested on 500 ulcer and 690 erosion images.

In contrast to these architectures, a two stage approach is proposed in [71]. RetinaNet [72] has been adapted for the initial detection stage where it receives an endoscopy image and predicts the classification scores and bounding boxes for the input image. Then they extract multiple fixed size patches of size  $160 \times 160$  from this image and pass those through a ResNet-18 [73] network where the final fully connected layer produces a binary classification for the detection of ulcers. This system has been tested with 4917 ulcer frames and 5007 normal frames and the model reaches 0.9469 ROC-AUC value.

Most recently, a two stream framework has been proposed in [74] where the authors extract features from two levels of the ResNet-50 [73] architecture, and they are combined using a relational network [75]. Fig. 10 illustrates this method. Specifically, the relational network allows this approach to map all possible relationships among the features extracted at the two levels. The resultant augmented feature vector is passed through an LSTM network and classification is performed using a fully connected layer. This framework has been evaluated on two public benchmarks, Kvasir [5] (with 8000 endoscopy images) and Nerthus [76] (with 2,552 colonoscopy images). In the Kvasir dataset this system was able to detect 8 abnormality classes with 98.4 % accuracy, and reaches 100 % accuracy for classifying the cleanliness of the

bowel on the Nerthus dataset. We note that this study exploits the hierarchical nature of the CNN to address the requirements in endoscopy image analysis. Top level kernels of a CNN capture local spatial features such as textures and contours in the input while the bottom level layers capture more semantic features, such as the overall representation of the image. This is because the local features are pooled together, hierarchically, when they flow through the CNN. Therefore, when extracting features from a CNN, top level layers carry spatially variant characteristics of the input while the bottom layers have spatially invariant features. The authors in [74] leverage this characteristic of CNNs for endoscopy image analysis in which, both existence of a particular distinctive pattern as well as its location is vital for diagnosis.

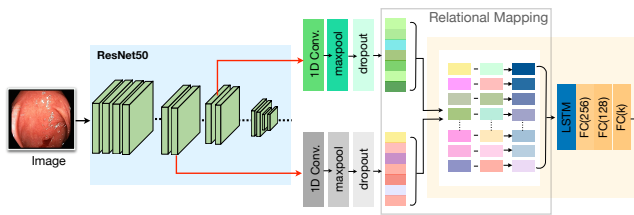


Fig. 10: The architecture proposed in [74] for abnormality detection in endoscopy data. Recreated from [74]

Similar to MRI image analysis, DCNNs have dominated endoscopy image abnormality detection. Furthermore, most methods utilise pre-trained feature extractors trained on large scale datasets with natural images leveraging the fact that endoscopy images are also captured with visible light. In contrast to binary supervised classification, methods such as [74] and [5] use multi-class classification (often trained using categorical cross-entropy loss) such that the model can detect normal and abnormal examples while recognising individual anomalies.

3) *Heart Sound Anomaly Detection*: In contrast to the MRI and endoscopy applications which use images, heart sound anomaly detection operates on a one-dimensional audio signal, and methods primarily use 1D CNNs and RNN architectures, however some pre-processing methods can be used to transform the audio signal to an image representation, allowing 2D CNNs to be used. An ensemble of VGG [77] networks is proposed in [78], where the authors first apply a Savitzky–Golay filter [79] to remove noise from the input signals. Then a series of 2D features, a spectrogram feature, a Mel Spectrogram and Mel Frequency Cepstral Coefficients (MFCCs), are extracted from the audio signal. These separate feature streams are passed through separate VGG networks and the final decision is made via majority voting. This method has also been evaluated on the PhysioNet/CinC 2016 dataset in a 10 fold cross validation setting and it reaches an accuracy of 89.81 %.

In [80] the authors leverage 497 feature values which are hand-crafted from 8 domains, including time domain features, higher-order statistics, signal energy, and frequency domain features. The extracted features are concatenated and passed through a 1D CNN with 3 convolutional layers followed by a

global average pooling layer and a dense layer with a sigmoid activation to perform the normal/abnormal classification. This system is evaluated on PhysioNet/CinC 2016 dataset and achieves an accuracy value of 86.8 %.

In contrast to the above stated approaches, frameworks that operate on raw audio signals are proposed in [81], [82]. Specifically, in [81] the authors augment the raw audio signal from the PhysioNet/CinC 2016 dataset by performing a Discrete Fourier Transform (DFT) and adding the variance and standard deviation of each data sample to the original audio. Then the recordings are segmented into S1 and S2 heart states using the algorithm of [83]. The segmented recordings are passed through an RNN to validate its normality. This framework achieves 80 % accuracy on the PhysioNet/CinC 2016 challenge. A similar approach utilising GRUs has been proposed in [82]. Similar to [81] the raw audio recordings were segmented to heart states using the algorithm of [83]. However, the authors in [82] skip the DFT based heart sound augmentation step utilised in [81]. The segmented audio is passed through a GRU network to generate the classification. The proposed framework has been validated for heart failure detection. The authors have acquired the heart failure data from patients in University-Town Hospital of Chongqing Medical University and the normal recordings were obtained from PhysioNet/CinC 2016 dataset (1286 randomly sampled normal recordings). In a 10-fold cross-validation setting the proposed model achieves an average accuracy of 98.82%. In this paper the authors have also tested the utilisation of an LSTM and Fully Convolutional Network (FCN) instead of a GRU network, however, these models have only been able to achieve 96.29 % and 94.65 %, respectively.

There has been a mixed response from researches regarding the need for heart sound segmentation prior to the abnormal heart sound detection. Heart sound segmentation has primarily been used due to the belief that features surrounding the S1 and S2 heart sound locations carry important information for detecting abnormalities. However, some argue that in the errors associated with this pre-processing step can be propagated to the abnormality detection module, and the model should be given the freedom to choose its own informative features [84]. In [84] the authors have conducted a comparative study investigating the importance of prior segmentation of heart sounds into heart states for abnormality detection. The authors have utilised the features extracted from the state-of-the-art the sound segmentation model [85] and trained a classifier to detect abnormalities using these features. For comparison, they also trained a separate 2D CNN model without segmentation which uses MFCC features as the inputs. The comparisons were conducted using the PhysioNet/CinC 2016 dataset and their evaluation indicates that a 2D CNN model without segmentation is capable of achieving superior results to a model that receives segmented inputs. In the 10-fold cross validation setting the unsegmented model achieves  $98.94 \pm 0.27$  % accuracy compared to  $98.49 \pm 0.13$  % for the segmented model. Utilising the SHAP model interpretations [86] the authors conclude that the unsegmented model has also focused on the regions of the audio wave that correspond to S1 and S2 locations, however, this model has the capacity to learn what

the informative features for the abnormality detection task are, compared to the restricted model inputs that are received by the segmented model.

Finally, Oh et. al [87] proposed a deep learned model called WaveNet to classify heart sounds into five categories, namely: normal, aortic stenosis, mitral valve prolapse, mitral stenosis, and mitral regurgitation. The architecture utilised in this study is illustrated in fig. 11. Specifically, inspired by [73] the authors proposed a residual block which is composed of 1D dilated convolutions to extract features from the raw audio signal. The architecture is composed of 6 such residual blocks and the features captured from those 6 blocks are aggregated into a single feature vector, which is subsequently passed through two 1D convolution layers and a two fully connected layers, prior to classification. This model is evaluated using an in house dataset which consists of 1000 PCG recording (200 per each category) and the model achieves an average accuracy of 97 % in a 10-fold cross validation setting.

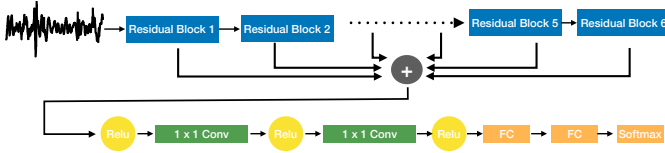


Fig. 11: The architecture proposed in [87] for abnormal heart sound detection. Recreated from [87]

As noted earlier, 1D-CNN networks and RNNs have been extensively applied for the abnormal heart sound detection. This is primarily due to the temporal nature of the signal where 1D-CNN networks can perform convolutions over the time axis and extract temporal features while the recurrent architectures can model the temporal evolution of the and generate better features for detecting the abnormalities. As discussed, there are only minor variations among the models and they have often utilised supervised learning to train the models. Furthermore, hand-crafted frequency domain features such as MFCCs are extensively applied within the heart sound anomaly detection domain as opposed to automatic feature learning. Finally, as observed in other application domains, supervised approaches are the most common methods for anomaly detection.

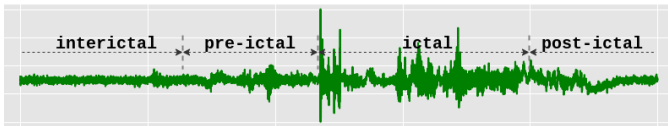


Fig. 12: Variations of the EEG recording before and after a seizure.

4) *Epileptic Seizure Prediction:* Fig. 12 illustrates how the four brain states: interictal, pre-ictal, ictal and post-ictal; are located in an EEG. The interictal state is the normal brain state of a subject, while the brain state before a seizure event is referred to as the pre-ictal state. The state in which the seizure occurs is denoted as ictal state, and after the seizure event, the brain shifts to the post-ictal state.

The seizure prediction problem can be viewed as an abnormality detection problem where machine learning models are trained to distinguish between the pre-ictal and interictal brain states, identifying when a particular subjects brain activity shifts from the normal interictal state to pre-ictal (abnormal state). As the pre-ictal state is the brain state before a seizure, this problem is termed seizure prediction.

We acknowledge that epileptic seizure prediction has several distinct characteristics compared to rest of the abnormality detection application domains that we discussed above, however, numerous studies have posed this task as an abnormality detection task [88], [89], [90], and hence we consider it here.

A key challenge in designing a generalised seizure prediction framework is the vast differences in the pre-ictal duration among subjects. This can vary from minutes to hours depending on the subject [91]. One of the notable attempts to perform patient independent seizure prediction is the framework of [92], where the authors propose a 2D CNN architecture trained on Short-term Fourier transform (STFT) features extracted from raw EEG signals. This framework has been validated on both the Freiburg intracranial EEG (iEEG) [93] and CHB-MIT scalp EEG (sEEG) datasets [94], and achieves approximately 81 % sensitivity in a leave-one-subject-out cross validation setting.

Despite this promising level of performance, the authors in [95], [96] identified significant performance variations in [92]. For example, the sensitivity drops to 33.3 % for some subjects. A multi-scale CNN architecture is proposed in [95] to address this limitation. The authors re-sample the original 400Hz iEEG dataset at 100Hz and STFT features are extracted from this down sampled signal. They extract STFT as 2D images for each EEG channel, resulting in 16 STFT images per data sample. The proposed multi-scale CNN is composed of 3 convolutional streams, each with different filter sizes ( $1 \times 1$ ,  $3 \times 3$  and  $5 \times 5$ ). The authors propose to capture features at different scales using these individual streams. These features are concatenated and passed through a fully connected layer to generate the relevant predictions. Their system is evaluated on 2016 Kaggle seizure prediction competition dataset [97] and the proposed system achieves a 87.85 % sensitivity where the lowest value per subject is only 79.65 %.

In contrast to this approach, a fine-tuning based method is proposed in [96]. The authors first train the model using a balanced dataset which consists of an equal amount of pre-ictal and interictal data. When the system is deployed, the authors propose to add a tunable processing layer which can be optimised depending on the patient requirements. This two stage framework is evaluated using the dataset proposed in [98] and the system achieve a mean sensitivity of 69%.

In contrast to these CNN based approaches, recurrent neural networks are leveraged in [89], [90]. Specifically, the authors in [89] utilised a 2 layer LSTM network trained on hand-crafted time domain, frequency domain, graph theory based (i.e clustering coefficients, diameter, radius, local efficiency, centrality, etc.), and correlation features. The system is evaluated using the CHB-MIT sEEG dataset and reaches 99.28% sensitivity for a 15 min pre-ictal period. Motivated by this approach, a bi-directional LSTM based architecture is given

in [90]. Similar to [89], a 2-layer LSTM is used with a bi-directional structure, however, in contrast to [89] it operates on the raw EEG signal. This framework has been validated using the Bonn University EEG database [99] and achieves an overall 89.2 % sensitivity score.

Compared to heart sound anomaly detection, most existing works in seizure prediction have utilised DCNN architectures. This is mainly due to the use of hand-crafted 2D image like features which are extracted jointly by considering all EEG electrodes. Once again, supervised learning methods are most prevalent and the architectures comprise standard deep learning methods.

A different approach is proposed by [100], who propose a GAN based method which is illustrated in Fig. 13. The generator of the GAN model is capable of synthesising realistic looking STFT images using a noise vector. The generated STFTs are passed through the discriminator which performs the real/fake validation. Once the generator is trained for the seizure prediction task, the authors adapt the discriminator network by adding two fully-connected layers such that it is trained to perform the normal/abnormal classification instead of real/fake classification. Therefore, the proposed system leverages the information in not only labeled EEG signals, but also the unlabeled synthesised samples in the training process. This system is validated using CHB-MIT sEEG, Freiburg iEEG, and EPILEPSIAE [101] datasets, and achieves AUC values of 77.68%, 75.47% and 65.05%, respectively.

We highlight that this approach deviates from the standard GAN model illustrated in Sec. II-B1, as in this model a secondary training process is used where the discriminator is fine-tuned to do normal/abnormal classification using supervised learning. Hence, like the autoencoder methods discussed for MRI anomaly detection in Section II-C1, this is not a completely unsupervised model. Rather this architecture is semi-supervised, where both labelled and unlabelled examples are used for model training [102].

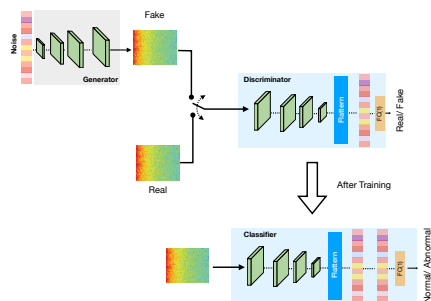


Fig. 13: The architecture proposed in [100] for epileptic seizure prediction. Recreated from [100]

### III. MODEL INTERPRETATION

Interpretability is one of the key challenges that modern deep learning methods face. Despite their tremendous success and often astonishingly precise predictions, the application of methods to real world diagnostic tasks is hindered as we are unsure how models reached their predictions. The

complexity of the deep learned models further contributes to this, as decisions are based upon hundreds of thousands of parameters, which are not human interpretable. Hence, interpretable machine learning has become an active area of research where black-box deep models are converted white-box models.

Fig. 14 illustrates a taxonomy of model interpretation methods, which is adopted in [103]. Model-agnostic interpretation methods are interpretation methods that are not limited to a specific architecture. In contrast, a model-specific interpretation method seeks to explain a single model.

Model interpretation methods can be further classified into local and global methods. Local methods try to reason regarding a particular prediction while global methods explore overall model behaviour by exploiting knowledge regarding the architecture, the training process and the data associated with training. The third class we consider is surrogate vs. visualization methods. In surrogate methods, a model with a simpler architecture (a surrogate model) is trained to mimic the behaviour of the original black box model. This is done with the intent that understanding the surrogate model's decision is simpler than the complex model. In contrast, visualisation methods use visual representations such as activation maps obtained from the black-box model to explain behaviour. Finally, as per [103] model interpretation techniques in the medical domain can be broadly categorised into attribution based and non-attribution based methods. Attribution-based methods seek to determine the contribution of an input feature to the generated classification. Non-attribution based methods investigate generating new methods to validate model behaviour for a given problem [103].

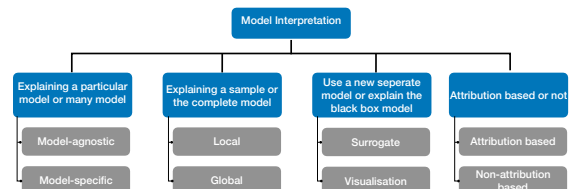


Fig. 14: Taxonomy of Model Interpretation methods.

The majority of existing literature on explainability of deep learned models in the medical domain considers attribution based methods. They leverage the model-agnostic plug and play nature of attribution based methods in their studies. The following paragraphs illustrate the most common model-agnostic interpretability methods that are used.

**Visualising Activation Maps:** This offers one of the simplest ways to understand what features lead to a certain model decision. As deep learning methods hierarchically encode features, the top layers of the model capture local features while later layers aggregate local features together to arrive at a decision. This concept is the foundation of Class Activation Maps (CAM) [104].

We can consider kernels in a convolution layer to be a set of filters which control information flow to subsequent layers, and at the final classification layer the positive features (emphasised features from the filters) are multiplied by learned



values to obtain a classification decision. Fig. 15 illustrates this concept. Hence the activation maps, or feature maps extracted at the final convolution layer, are multiplied by the associated weights and they are aggregated to generate the final activation map of the predicted class. The resultant map is up-sampled such that it can be superimposed on the input image. This can reveal what regions/characteristics of the input are highly activated and pass information to the classifier. Such a technique can be applied for tasks such as CNN based MRI tumor detection to identify whether the features from the tumor region are actually contributing to the classification, or if the model is acting upon noisy features from elsewhere in the sample.

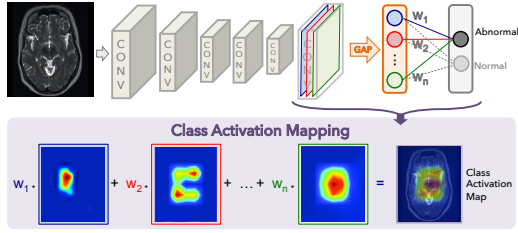


Fig. 15: Illustration of the process of creating class activation maps. Recreated from [104]

One of the drawbacks of the CAM generation process is that the technique is constrained to architectures where global average pooling is performed over convolutional maps immediately before prediction. This is a requirement for utilising the weighted liner sum of the activations to generate the final convolution map.

The Grad-Cam [105] method addresses this shortcoming, and uses the gradient information flowing into the last convolutional layer of the network to understand how each pixel contributes to the decision. Let the  $k^{th}$  feature map of the final convolution layer of size  $u \times v$  be denoted by  $A^k$ . Then the gradients of the score for class  $c$ ,  $y^c$ , are computed with respect to feature maps  $A^k$ , and averaged across  $u \times v$  such that,

$$\alpha^{c,k} = \frac{1}{uv} \sum_{i \in u} \sum_{j \in v} \frac{\partial y^c}{\partial A_{i,j}^k} \quad (23)$$

Then to generate the final activation map across all  $k$  feature maps a weighted combination of activation maps is computed. The resultant feature map is passed through the ReLU activation to set negative values to zero.

$$L_{Grad-Cam} = \text{ReLU}(\sum_k \alpha^{c,k} A^k) \quad (24)$$

Grad-Cam however does not handle instances where multiple occurrences of the same object are in the input image, and in such instances it fails to properly localise the multiple instances [106].

**Local Interpretable Model-agnostic Explanations (LIME):** As the name implies, LIME is a local interpretation method. It tries to interpret a model's behaviour when presented with different inputs by understanding how predictions change with perturbations of the input data. Fig.

16 illustrates the concept behind LIME. First, the input is divided to a series of interpretable components where some portion of the input is masked out. Then each perturbed sample is passed through the model to get the probability of a particular class and the components of the image with the highest weights are returned as the explanation.

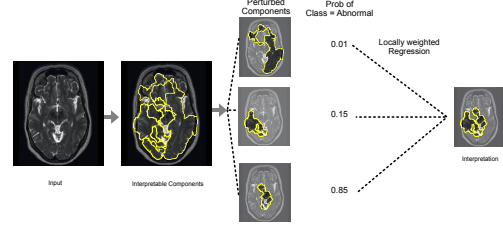


Fig. 16: Illustration of Local Interpretable Model-agnostic Explanations method. Recreated from [107]

One of the key limitations of LIME is that when sampling the data points for interpretable components, it can sample unrealistic data points. Furthermore, as the interpretations are biased towards these data points the generated explanations can be unstable such that two components in close proximity can lead to very different explanations.

**SHapley Additive exPlanations (SHAP):** The inspiration for SHAP [108] comes from game theory. Specifically, we define individual feature values (or groups of feature values) as a 'player' in the game and the contribution of each player to the game is measured by adding and removing the players. Let the input  $x$  be composed of  $N$  features,  $\omega_i$ s, where  $x = [\omega_1, \omega_2, \dots, \omega_N]$ , and  $M$  is the maximum number of coalitions (or feature combinations) that can be generated using  $N$  features. Then the model is queried with different feature coalitions where some feature values are present and some are absent. (eg.  $x_1 = [\omega_1]$ ,  $x_2 = [\omega_2]$ ,  $x_{1,2} = [\omega_1, \omega_2]$ ,  $x_{1,3} = [\omega_1, \omega_3]$ , ...). This allows the identification of which features contribute to a certain prediction and which do not.

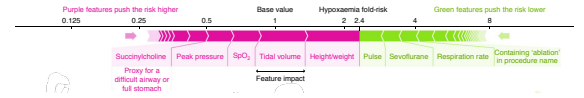


Fig. 17: Illustration of SHapley Additive exPlanations which explore how different features contribute to the risk of hypoxaemia. Image taken from [109]

Fig. 17 illustrates how different features such as height/weight, respiration rate and pulse affect the risk of hypoxaemia in the next five minutes. Features shown in purple increase the risk while green features reduce the risk.

In [86] the author suggests that SHAP is one of the few explanation method with a strong theoretical basis and the only method that currently exists to deliver a full interpretation. However, it is computationally expensive to calculate SHAP values as we have to consider all possible combinations of the features.

**Interpretation of Medical Anomaly Detection Methods:** In addition to the above stated commonly used interpretation methods we acknowledge the methods DeepLIFT



(Deep Learning Important Features) [110], DeepTaylor [111], Guided backpropagation (GBP) [112] and Integrated Gradients [113], that are also proposed to explain the black-box deep learning model.

In [114] the authors attempt to explain the features learned by a CNN which they proposed for automated grading of brain tumors from MRI images using GradCAM and GBP techniques. In [115] 30 CNN models were trained for melanoma detection using skin images and the authors interpret the model features using SHAP and GradCAM. They illustrate that models occasionally focus on features that were irrelevant for diagnosis.

In recent studies [116], [117] GradCAM, GBP, CAM, DeepLIFT and IG have been utilised to explain chest X-ray based COVID-19 detection of a deep learned model. Most recently, SHAP interpretations are used to illustrate that a heart sound anomaly detection methods can automatically learn to focus on S1 and S2 sounds without the need to provide segmented inputs [84].

In contrast to the attribution based approaches illustrated earlier, non-attribution based methods use concepts like attention maps, and expert knowledge to interpret model decisions. However, these methods are specific to a particular problem. For instance, in [118] attention is used to map the relationships between medical images and corresponding diagnostic reports. This mechanism uncovers the mapping between images and the diagnostic reports. A textual justification for a breast mass classification task is proposed in [119]. The proposed justification model receives visual features and embeddings from the classifier and generates a diagnostic sentence explaining the model classification. In [120] a method for generating understandable explanations utilising a set of explainable rules is presented. In this approach, a set of anatomical features are defined based on segmentation and anatomical regularities (i.e. set of pre-defined rules), and the feature importance is evaluated based on perturbation.

Considering the above discussion it is clear that the selection of the interpretation method depends on several factors:

- 1) whether a global level model interpretation method is required, or whether it is sufficient to generate local (example level) interpretations;
- 2) the end users expertise level with regards to understanding the resultant explanations; and
- 3) whether the application domain has time constraints, i.e. do the interpretations need to be generated in real-time?

Interpretable machine learning is an active area of research and the medical domain is a good test bed to evaluate the proposed methods. Better understanding regarding the black-box model decisions would not only build trust among the medical practitioners regarding machine learning algorithms, but also would help the machine learning researchers to understand the limitations of model architectures and help to design better models. However, as illustrated earlier, different interpretation approaches have different strengths and limitations, and designing optimal interpretation strategies is an open research problem for future exploration.

#### IV. CHALLENGES AND OPEN RESEARCH QUESTIONS

In this section, we outline limitations of existing deep medical anomaly detection techniques as well as various open research questions, and highlight future research directions.

##### A. Lack of Interpretability

As illustrated in Sec. III, attribution based methods have been popular among researchers in the medical domain for deep model interpretation due to their model agnostic plug-and-play nature. However, the end-user of the given particular medical application (i.e. the clinician) should be considered when selecting one interpretation method over another. Although popular, methodologies such as GradCAM, LIME, and GBP are not specifically developed to address explainability in the medical domain, and while they are informative for machine learning practitioners, they may be of much less use for a clinicians. Therefore, more studies such as [121], [122] should be conducted using expert clinicians to rate the explanations across different application domains. Such illustrations would evaluate the applicability and the limitations of model-agnostic interpretation methods. Hybrid techniques such as Human-in-the-Loop learning techniques could be utilised to design interpretable diagnostic models where clinical experts could refine deep model decisions to mimic their own decision making process.

Furthermore, we observe a lack of model-agnostic methods to interpret multi-modal deep methodologies. Such methods have increased complexity in that the decision depends on multiple input feature streams, requiring more sophisticated strategies to interpret behaviour.

Finally, we present Reinforcement Learning (RL) as a possible future direction to generate explainable decisions [123], [124]. In RL the autonomous agent's behaviour is governed by a 'reward function', and the agent tries to maximise this reward. As such the agent utilises exploration to detect anomalies and improve its detection process across many iterations. The exploration process that the agent utilised to detect the anomalies could illustrate the intuition behind its behaviour.

##### B. Causality and Uncertainty

Causal identification is crucial characteristic that most existing deep medical anomaly detection methods lack. Causality is often confused with association [125], [126], [127]. For instance, if  $X$  and  $Y$  are associated (dependent) it only implies that there is a dependency between the two factors. The association does not imply that  $X$  is causing  $Y$ . Association can arise between variables in the presence and absence of a causal relationship. If both  $X$  and  $Y$  have a common cause they both can show associative relationships without causality [127].

In medical diagnosis the doctor tries to explain the cause of the patient's symptoms, which is causal identification. However, in most existing deep learned approaches the diagnosis is purely associative. Methods try to associate the patient's symptoms with a particular disease category without trying

to uncover what is actually causing these symptoms (and whether this disease is the only cause of these symptoms) [125]. As such, causality estimation is a crucial area that requires additional focus from the research community. For instance, in epilepsy prediction if the the brain regions that are actually causing the seizures can be identified then epileptologists can surgically treat that specific region to address the root cause of the patient’s seizures. Existing approaches for causality estimation in deep learning studies include causal graph structures [128], algorithmic information theory based approaches [129], [130], and Causal Bayesian Networks [131], [132]; however, these methods are seldom applied in medical abnormality detection.

Uncertainty estimation is another characteristic that most current state-of-the-art anomaly detection algorithms lack. Such methods quantitatively estimate how a small change in input parameters affects model predictions. This can be indicative of model confidence. For instance, Bayesian Deep Learning [133] could be used to generate probabilistic scores, where the model parameters are approximated through a variational method, generating uncertainty information regarding the model weights, such that one can observe the uncertainty of the model outputs. We would like to acknowledge the work of Leibig et. al [134] where they illustrate how the computed measure of uncertainty can be utilised to refer a subset of difficult cases for further inspection. Furthermore, Bayesian uncertainty estimation has been applied for estimating the quality of medical image segmentation [135], [136], [137], and sclerosis lesion detection [138]. We believe further extensive investigation will allow rapid application of uncertainty estimation measure in the medical anomaly detection algorithms.

### C. Lack of Generalisation

Another limitation that we observe in present research is the lack of generalisation to different operating conditions. For instance, [139] observed an abnormal heart sound detection model that achieves more than 99 % accuracy drop to 52.27 % when presented with an unseen dataset. Such performance instability significantly hinders the applicability of the deep learned models in real world life-or-death medical applications. One of the major reasons for such specificity of the models is data scarcity in the medical domain. Even though the number of datasets that are publicly available continues to increase, there are still a limited number of data samples available (compared to large scale datasets such as ImageNet). Most datasets are also highly curated, collected in controlled environments and restricted settings that do not capture the real data distribution. For instance, in Fig. 18 we visualise the t-SNE plot generated for the model in [139] using 2,000 randomly chosen samples (which contain both normal and abnormal samples) from PhysioNet/CinC 2016 dataset. This dataset is composed of 6 sub datasets (denoted a-f in the figure), collected from different devices (Welch Allyn Meditron, 3M Littmann, and ABES Electronic stethoscope), different capture environments (Lab setting and hospitals), varying age groups, etc.

As illustrated in Figure 18, the samples of each subset are somewhat grouped together while the samples from different

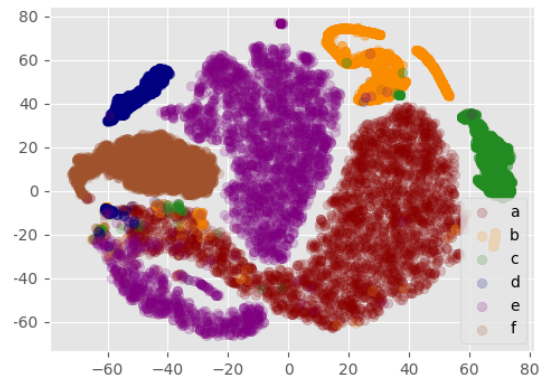


Fig. 18: Visualisation of t-SNE plots for different domains in PhysioNet/CinC 2016 dataset abnormal heart sound detection challenge. Image taken from [139]

subsets are distributed across the embedding space. This example clearly illustrates the challenges associated with medical anomaly detection as the acquired samples may not optimally capture the population characteristics. If a diagnostic model is trained only on a particular sub-set of this dataset it would generate erroneous detections on another. Therefore, large scale datasets which capture the diverse nature of the full population are required.

While large scale datasets akin to ImageNet are ideal, it is very expensive and sometimes not practically feasible to collect large scale annotated datasets in medical diagnostic research [140], [141]. Therefore, meta-learning and domain adaptation approaches could be of use when annotated examples are scarce. In particular, meta-learning, which is a sub-field of transfer learning, focuses on how a model trained for a particular task can be quickly adapted to a novel task. Hence, the learned knowledge is shared between the two tasks. Meta-learning is an emerging technique in the medical domain [142], [141], [140], and could be extensively utilised to train large scale models using limited data samples. In contrast to meta-learning, domain adaptation focuses on how a generalised model trained for the same task can be adapted to a particular sub-domain (such as subsets a-f in Fig. 18) [139]. Such approaches can also be utilised to attain generalisation in medical anomaly detection such that a model trained on a specific domain can be adapted to other domains using few labeled examples.

### D. Handling Data Imbalance and Unlabelled Data

The majority of existing public medical abnormality detection benchmarks are highly imbalanced in terms of normal and abnormal sample counts. In most scenarios it is comparatively easy to obtain normal samples compared to anomalous samples, yielding imbalanced datasets. This typically becomes an issue in supervised training as the model becomes more sensitive to the loss arising from the majority class, compared to classes with fewer examples. The most common approaches to address class imbalance in medical anomaly detection has been data re-sampling (under or over sampling) and cost

sensitive training where more weight in the loss is assigned to the minority class [143].

However, data augmentation strategies such as using GANs have recently emerged which are capable of generating synthetic data for training, and they are favoured over traditional methods for handling data imbalance [144], [145]. For instance, in [146], [144], [145] the generator is used to synthesise realistic-looking minority class samples, thereby balancing the class distribution and avoiding overfitting. Despite their superior results compared to traditional methods, generating realistic looking data samples is an open research problem [145]. Further research is required to improve the quality of the synthesised samples and to determine effective GAN learning strategies that can better adapt to novelties in the abnormal (which is typically the minority) class.

Another interesting research direction for investigation is methods to handle unlabelled data. In most scenarios it is cheaper to obtain unlabelled data compared to labelled samples. Hence, if the training mechanism can leverage information in unlabelled samples, it could be highly beneficial. The sub-field of semi-supervised learning addresses this situation and GANs have also demonstrated tremendous success in a semi-supervised setting [102] where the trained discriminator is adapted to perform the normal/abnormal classification task, instead of real/fake validation [100]. However, we observe that deep medical anomaly detection methods rarely utilise semi-supervised learning strategies. Hence, further investigation should be carried out to introduce such strategies into the medical domain.

In addition to semi-supervised learning, self-supervised learning is another new research direction with significant potential. In contrast to semi-supervised learning, self-supervised learning considers learning from internal cues. In particular, it uses preliminary tasks such as context prediction [147], colorization [148], and design a jigsaw puzzle game [149] to pre-train the model such that it learns about the data distribution. Most importantly, these pretext tasks do not require labelled data and the objectives are designed to generate labels automatically. Then, the learned knowledge is transferred to different downstream tasks such as image classification, object detection, and action recognition.

Recent works have investigated self-supervised learning for anomaly detection. For instance, in [150] the authors investigate the objective of predicting the indices of randomly permuted video frames as the self-supervised objective. The authors show that implicitly reasoning about the relative positions of the objects and their motions, which is beneficial to detect abnormal behaviour. However, we observe that self-supervised learning has not yet emerged into the medical anomaly detection domain. We observe the potential of utilising pretext tasks such as medical image segmentation, artificially synthesising rotated images as self-supervised objectives in this field. Therefore, further investigations can be carried out to assess the viability of such techniques.

## V. CONCLUSION

In this survey paper, we have discussed various approaches across deep learning-based medical anomaly detection. In

particular, we have outlined different data capture settings across different medical applications, numerous deep learning architectures that have been motivated due to these different data types and problem specifications, and various learning strategies that have been applied. This structured analysis of deep medical anomaly detection research methodologies enabled comparing and contrasting existing state-of-the-art techniques despite their application differences. Moreover, we provided a comprehensive overview of deep model interpretation strategies, outlining the strengths and weaknesses of those interpretation mechanisms. As concluding remarks, we outlined key limitations of existing deep medical anomaly detection techniques and proposed possible future research directions for further investigation.

## REFERENCES

- [1] T. Fernando, S. Denman, D. Ahmedt-Aristizabal, S. Sridharan, K. R. Laurens, P. Johnston, and C. Fookes, "Neural memory plasticity for medical anomaly detection," *Neural Networks*, 2020.
- [2] S. Thudumu, P. Branch, J. Jin, and J. J. Singh, "A comprehensive survey of anomaly detection techniques for high dimensional big data," *Journal of Big Data*, vol. 7, no. 1, pp. 1–30, 2020.
- [3] R. Chalapathy and S. Chawla, "Deep learning for anomaly detection: A survey," *arXiv preprint arXiv:1901.03407*, 2019.
- [4] Z. Zhao, S. Cerf, R. Birke, B. Robu, S. Bouchenak, S. B. Mokhtar, and L. Y. Chen, "Robust anomaly detection on unreliable data," in *2019 49th Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN)*. IEEE, 2019, pp. 630–637.
- [5] K. Pogorelov, K. R. Randel, C. Griwodz, S. L. Eskeland, T. de Lange, D. Johansen, C. Spampinato, D.-T. Dang-Nguyen, M. Lux, P. T. Schmidt *et al.*, "Kvasir: A multi-class image dataset for computer aided gastrointestinal disease detection," in *Proceedings of the 8th ACM on Multimedia Systems Conference*, 2017, pp. 164–169.
- [6] G. D. Clifford, C. Liu, B. Moody, D. Springer, I. Silva, Q. Li, and R. G. Mark, "Classification of normal/abnormal heart sound recordings: The physionet/computing in cardiology challenge 2016," in *2016 Computing in Cardiology Conference (CinC)*. IEEE, 2016, pp. 609–612.
- [7] N. Alahmadi, S. A. Evdokimov, Y. J. Kropotov, A. M. Müller, and L. Jäncke, "Different resting state eeg features in children from switzerland and saudi arabia," *Frontiers in human neuroscience*, vol. 10, p. 559, 2016.
- [8] N. Gönitz, M. Kloft, K. Rieck, and U. Brefeld, "Toward supervised anomaly detection," *Journal of Artificial Intelligence Research*, vol. 46, pp. 235–262, 2013.
- [9] A. L. Beam and I. S. Kohane, "Big data and machine learning in health care," *Jama*, vol. 319, no. 13, pp. 1317–1318, 2018.
- [10] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
- [11] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *arXiv preprint arXiv:1412.3555*, 2014.
- [12] K. Rasheed, A. Qayyum, J. Qadir, S. Sivathamboo, P. Kwan, L. Kuhlmann, T. O'Brien, and A. Razi, "Machine learning for predicting epileptic seizures using eeg signals: A review," *arXiv preprint arXiv:2002.01925*, 2020.
- [13] S. Li, F. Li, S. Tang, and W. Xiong, "A review of computer-aided heart sound detection techniques," *BioMed Research International*, vol. 2020, 2020.
- [14] W. Du, N. Rao, D. Liu, H. Jiang, C. Luo, Z. Li, T. Gan, and B. Zeng, "Review on the applications of deep learning in the analysis of gastrointestinal endoscopy images," *IEEE Access*, vol. 7, pp. 142 053–142 069, 2019.
- [15] A. S. Lundervold and A. Lundervold, "An overview of deep learning in medical imaging focusing on mri," *Zeitschrift für Medizinische Physik*, vol. 29, no. 2, pp. 102–127, 2019.
- [16] S. Soffer, E. Klang, O. Shimon, N. Nachmias, R. Eliakim, S. Ben-Horin, U. Kopylov, and Y. Barash, "Deep learning for wireless capsule endoscopy: a systematic review and meta-analysis," *Gastrointestinal Endoscopy*, 2020.

- [17] Z. Ebrahimi, M. Loni, M. Daneshalab, and A. Gharehbaghi, "A review on deep learning methods for ecg arrhythmia classification," *Expert Systems with Applications: X*, p. 100033, 2020.
- [18] D. H. Ballard, "Modular learning in neural networks," in *AAAI*, 1987, pp. 279–284.
- [19] D. Charte, F. Charte, S. García, M. J. del Jesus, and F. Herrera, "A practical tutorial on autoencoders for nonlinear feature fusion: Taxonomy, models, software and guidelines," *Information Fusion*, vol. 44, pp. 78–96, 2018.
- [20] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol, "Extracting and composing robust features with denoising autoencoders," in *Proceedings of the 25th international conference on Machine learning*, 2008, pp. 1096–1103.
- [21] J. Cawton, I. Kyriazakis, T. Plötz, and J. Bacardit, "A combined deep learning gru-autoencoder for the early detection of respiratory disease in pigs using multiple environmental sensors," *Sensors*, vol. 18, no. 8, p. 2521, 2018.
- [22] K. Wang, Y. Zhao, Q. Xiong, M. Fan, G. Sun, L. Ma, and T. Liu, "Research on healthy anomaly detection model based on deep learning from multiple time-series physiological signals," *Scientific Programming*, vol. 2016, 2016.
- [23] D. Sato, S. Hanaoka, Y. Nomura, T. Takenaga, S. Miki, T. Yoshikawa, N. Hayashi, and O. Abe, "A primitive study on unsupervised anomaly detection with an autoencoder in emergency head ct volumes," in *Medical Imaging 2018: Computer-Aided Diagnosis*, vol. 10575, International Society for Optics and Photonics, 2018, p. 105751P.
- [24] Y. Lu and P. Xu, "Anomaly detection for skin disease images using variational autoencoder," *arXiv preprint arXiv:1807.01349*, 2018.
- [25] D. Zimmerer, S. A. Kohl, J. Petersen, F. Isensee, and K. H. Maier-Hein, "Context-encoding variational autoencoder for unsupervised anomaly detection," *arXiv preprint arXiv:1812.05941*, 2018.
- [26] H. Uzunova, S. Schultz, H. Handels, and J. Ehrhardt, "Unsupervised pathology detection in medical images using conditional variational autoencoders," *International journal of computer assisted radiology and surgery*, vol. 14, no. 3, pp. 451–461, 2019.
- [27] D. Saxena and J. Cao, "Generative adversarial networks (gans): Challenges, solutions, and future directions," *arXiv preprint arXiv:2005.00065*, 2020.
- [28] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672–2680.
- [29] T. Fernando, S. Denman, S. Sridharan, and C. Fookes, "Task specific visual saliency prediction with memory augmented conditional generative adversarial networks," in *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2018, pp. 1539–1548.
- [30] —, "Gd-gan: Generative adversarial networks for trajectory prediction and group detection in crowds," in *Asian Conference on Computer Vision*. Springer, 2018, pp. 314–330.
- [31] T. Schlegl, P. Seeböck, S. M. Waldstein, G. Langs, and U. Schmidt-Erfurth, "f-anogan: Fast unsupervised anomaly detection with generative adversarial networks," *Medical image analysis*, vol. 54, pp. 30–44, 2019.
- [32] Z. Zhang and M. Sabuncu, "Generalized cross entropy loss for training deep neural networks with noisy labels," in *Advances in neural information processing systems*, 2018, pp. 8778–8788.
- [33] U. Schmidt-Erfurth, A. Sadeghipour, B. S. Gerendas, S. M. Waldstein, and H. Bogunović, "Artificial intelligence in retina," *Progress in retinal and eye research*, vol. 67, pp. 1–29, 2018.
- [34] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun, "Dermatologist-level classification of skin cancer with deep neural networks," *nature*, vol. 542, no. 7639, pp. 115–118, 2017.
- [35] J. Turner, A. Page, T. Mohsenin, and T. Oates, "Deep belief networks used on high resolution multichannel electroencephalography data for seizure detection," *arXiv preprint arXiv:1708.08430*, 2017.
- [36] P. K. Jawanpuria, M. Lapin, M. Hein, and B. Schiele, "Efficient output kernel learning for multiple tasks," in *Advances in neural information processing systems*, 2015, pp. 1189–1197.
- [37] P. Kisilev, E. Sason, E. Barkan, and S. Hashoul, "Medical image description using multi-task-loss cnn," in *Deep Learning and Data Labeling for Medical Applications*. Springer, 2016, pp. 121–129.
- [38] R. Williams, "Gradient-based learning algorithm for recurrent networks," *Back-propagation: theory, architectures and applications*, 1995.
- [39] S. Hochreiter, "The vanishing gradient problem during learning recurrent neural nets and problem solutions," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 6, no. 02, pp. 107–116, 1998.
- [40] Y. Bengio, P. Simard, and P. Frasconi, "Learning long-term dependencies with gradient descent is difficult," *IEEE transactions on neural networks*, vol. 5, no. 2, pp. 157–166, 1994.
- [41] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," *arXiv preprint arXiv:1406.1078*, 2014.
- [42] T. Fernando, S. Denman, A. McFadyen, S. Sridharan, and C. Fookes, "Tree memory networks for modelling long-term temporal dependencies," *Neurocomputing*, vol. 304, pp. 64–81, 2018.
- [43] T. Fernando, S. Denman, S. Sridharan, and C. Fookes, "Learning temporal strategic relationships using generative adversarial imitation learning," in *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*. International Foundation for Autonomous Agents and Multiagent Systems, 2018, pp. 113–121.
- [44] —, "Memory augmented deep generative models for forecasting the next shot location in tennis," *IEEE Transactions on Knowledge and Data Engineering*, 2019.
- [45] H. Gammulle, S. Denman, S. Sridharan, and C. Fookes, "Forecasting future action sequences with neural memory networks," *British Machine Vision Conference (BMVC)*, 2019.
- [46] Y. Ma and J. C. Principe, "A taxonomy for neural memory networks," *IEEE transactions on neural networks and learning systems*, 2019.
- [47] T. Munkhdalai and H. Yu, "Neural semantic encoders," in *Proceedings of the conference. Association for Computational Linguistics. Meeting*, vol. 1. NIH Public Access, 2017, p. 397.
- [48] A. N. Jagannatha and H. Yu, "Bidirectional rnn for medical event detection in electronic health records," in *Proceedings of the conference. Association for Computational Linguistics. North American Chapter. Meeting*, vol. 2016. NIH Public Access, 2016, p. 473.
- [49] H. Yang and H. Gao, "Toward sustainable virtualized healthcare: extracting medical entities from chinese online health consultations using deep neural networks," *Sustainability*, vol. 10, no. 9, p. 3292, 2018.
- [50] S. Latif, M. Usman, R. Rana, and J. Qadir, "Phonocardiographic sensing using deep learning for abnormal heartbeat detection," *IEEE Sensors Journal*, vol. 18, no. 22, pp. 9393–9400, 2018.
- [51] D. Ahmedt-Aristizabal, T. Fernando, S. Denman, L. Petersson, M. J. Aburn, and C. Fookes, "Neural memory networks for robust classification of seizure type," *International Conferences of the IEEE Engineering in Medicine and Biology Society*, 2020.
- [52] Y. Yoo, L. Y. Tang, T. Brosch, D. K. Li, S. Kolind, I. Vavasour, A. Rauscher, A. L. MacKay, A. Traboulsee, and R. C. Tam, "Deep learning of joint myelin and t1w mri features in normal-appearing brain tissue to distinguish between multiple sclerosis patients and healthy controls," *NeuroImage: Clinical*, vol. 17, pp. 169–178, 2018.
- [53] G. E. Hinton, "Deep belief networks," *Scholarpedia*, vol. 4, no. 5, p. 5947, 2009.
- [54] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.
- [55] D. Lu, K. Popuri, G. W. Ding, R. Balachandar, and M. F. Beg, "Multimodal and multiscale deep neural networks for the early diagnosis of alzheimer's disease using structural mr and fdg-pet images," *Scientific reports*, vol. 8, no. 1, pp. 1–13, 2018.
- [56] C. R. Jack Jr, M. A. Bernstein, N. C. Fox, P. Thompson, G. Alexander, D. Harvey, B. Borowski, P. J. Britson, J. L. Whitwell, C. Ward *et al.*, "The alzheimer's disease neuroimaging initiative (adni): Mri methods," *Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine*, vol. 27, no. 4, pp. 685–691, 2008.
- [57] M. H. Le, J. Chen, L. Wang, Z. Wang, W. Liu, K.-T. T. Cheng, and X. Yang, "Automated diagnosis of prostate cancer in multi-parametric mri based on multimodal convolutional neural networks," *Physics in Medicine & Biology*, vol. 62, no. 16, p. 6497, 2017.
- [58] J. Islam and Y. Zhang, "Brain mri analysis for alzheimer's disease diagnosis using an ensemble system of deep convolutional neural networks," *Brain informatics*, vol. 5, no. 2, p. 2, 2018.
- [59] D. Marcus, T. Wang *et al.*, "Oasis: Cross-sectional mri data in young, middle aged, nondemented, and demented older adults," *Journal of Cognitive Neuroscience*.
- [60] M. Shehata, F. Khalifa, A. Soliman, M. Ghazal, F. Taher, M. Abou El-Ghar, A. C. Dwyer, G. Gimel'farb, R. S. Keynton, and A. El-Baz, "Computer-aided diagnostic system for early detection of acute renal transplant rejection using diffusion-weighted mri," *IEEE Transactions on Biomedical Engineering*, vol. 66, no. 2, pp. 539–552, 2018.

- [61] L.-L. Zeng, H. Wang, P. Hu, B. Yang, W. Pu, H. Shen, X. Chen, Z. Liu, H. Yin, Q. Tan *et al.*, "Multi-site diagnostic classification of schizophrenia using discriminant deep learning with functional connectivity mri," *EBioMedicine*, vol. 30, pp. 74–85, 2018.
- [62] Z. Han, B. Wei, S. Leung, I. B. Nachum, D. Laidley, and S. Li, "Automated pathogenesis-based diagnosis of lumbar neural foraminal stenosis via deep multiscale multitask learning," *Neuroinformatics*, vol. 16, no. 3-4, pp. 325–337, 2018.
- [63] J. Cheng, W. Huang, S. Cao, R. Yang, W. Yang, Z. Yun, Z. Wang, and Q. Feng, "Enhanced performance of brain tumor classification via tumor region augmentation and partition," *PloS one*, vol. 10, no. 10, p. e0140381, 2015.
- [64] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A very deep convolutional networks for large-scale image recognition hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.
- [65] E. Klang, Y. Barash, R. Y. Margalit, S. Soffer, O. Shimon, A. Albsheh, S. Ben-Horin, M. M. Amitai, R. Eliakim, and U. Kopylov, "Deep learning algorithms for automated detection of crohn's disease ulcers by video capsule endoscopy," *Gastrointestinal Endoscopy*, vol. 91, no. 3, pp. 606–613, 2020.
- [66] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1251–1258.
- [67] H. Alaskar, A. Hussain, N. Al-Aseem, P. Liatsis, and D. Al-Jumeily, "Application of convolutional neural networks for automated ulcer detection in wireless capsule endoscopy images," *Sensors*, vol. 19, no. 6, p. 1265, 2019.
- [68] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.
- [69] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.
- [70] S. Fan, L. Xu, Y. Fan, K. Wei, and L. Li, "Computer-aided detection of small intestinal ulcer and erosion in wireless capsule endoscopy images," *Physics in Medicine & Biology*, vol. 63, no. 16, p. 165001, 2018.
- [71] S. Wang, Y. Xing, L. Zhang, H. Gao, and H. Zhang, "A systematic evaluation and optimization of automatic detection of ulcers in wireless capsule endoscopy on a large dataset using deep convolutional neural networks," *Physics in Medicine & Biology*, vol. 64, no. 23, p. 235014, 2019.
- [72] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980–2988.
- [73] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
- [74] H. Gammulle, S. Denman, S. Sridharan, and C. Fookes, "Two-stream deep feature modelling for automated video endoscopy data analysis," *International Conference on Medical Image Computing and Computer Assisted Intervention*, 2020.
- [75] A. Santoro, D. Raposo, D. G. Barrett, M. Malinowski, R. Pascanu, P. Battaglia, and T. Lillicrap, "A simple neural network module for relational reasoning," in *Advances in neural information processing systems*, 2017, pp. 4967–4976.
- [76] K. Pogorelov, K. R. Randel, T. de Lange, S. L. Eskeland, C. Griwodz, D. Johansen, C. Spampinato, M. Taschwer, M. Lux, P. T. Schmidt *et al.*, "Nerthus: A bowel preparation quality video dataset," in *Proceedings of the 8th ACM on Multimedia Systems Conference*, 2017, pp. 170–174.
- [77] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.
- [78] J. M.-T. Wu, M.-H. Tsai, Y. Z. Huang, S. H. Islam, M. M. Hassan, A. Alelaiwi, and G. Fortino, "Applying an ensemble convolutional neural network with savitzky-golay filter to construct a phonocardiogram prediction model," *Applied Soft Computing*, vol. 78, pp. 29–40, 2019.
- [79] A. Savitzky and M. J. Golay, "Smoothing and differentiation of data by simplified least squares procedures," *Analytical chemistry*, vol. 36, no. 8, pp. 1627–1639, 1964.
- [80] F. Li, H. Tang, S. Shang, K. Mathiak, and F. Cong, "Classification of heart sounds using convolutional neural network," *Applied Sciences*, vol. 10, no. 11, p. 3956, 2020.
- [81] T.-c. I. Yang and H. Hsieh, "Classification of acoustic physiological signals based on deep learning neural networks with augmented features," in *2016 Computing in Cardiology Conference (CinC)*. IEEE, 2016, pp. 569–572.
- [82] S. Gao, Y. Zheng, and X. Guo, "Gated recurrent unit-based heart sound analysis for heart failure screening," *BioMedical Engineering OnLine*, vol. 19, no. 1, p. 3, 2020.
- [83] D. B. Springer, L. Tarassenko, and G. D. Clifford, "Logistic regression-hsnn-based heart sound segmentation," *IEEE Transactions on Biomedical Engineering*, vol. 63, no. 4, pp. 822–832, 2015.
- [84] T. Dissanayake, T. Fernando, S. Denman, S. Sridharan, H. Ghaemmaghami, and C. Fookes, "A robust interpretable deep learning classifier for heart anomaly detection without segmentation," *IEEE Journal of Biomedical and Health Informatics*, 2020.
- [85] T. Fernando, H. Ghaemmaghami, S. Denman, S. Sridharan, N. Hussain, and C. Fookes, "Heart sound segmentation using bidirectional lstms with attention," *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 6, pp. 1601–1609, 2019.
- [86] C. Molnar, *Interpretable Machine Learning*. Lulu.com, 2020.
- [87] S. L. Oh, V. Jahmunah, C. P. Ooi, R.-S. Tan, E. J. Ciaccio, T. Yamakawa, M. Tanabe, M. Kobayashi, and U. R. Acharya, "Classification of heart sound signals using a novel deep wavenet model," *Computer Methods and Programs in Biomedicine*, p. 105604, 2020.
- [88] K. Gadhouri, J.-M. Lina, and J. Gotman, "Discriminating preictal and interictal states in patients with temporal lobe epilepsy using wavelet analysis of intracerebral eeg," *Clinical neurophysiology*, vol. 123, no. 10, pp. 1906–1916, 2012.
- [89] K. M. Tsiouris, V. C. Pezoulas, M. Zervakis, S. Konitsiotis, D. D. Koutsouris, and D. I. Fotiadis, "A long short-term memory deep learning network for the prediction of epileptic seizures using eeg signals," *Computers in biology and medicine*, vol. 99, pp. 24–37, 2018.
- [90] D. Thara, B. PremaSudha, and F. Xiong, "Epileptic seizure detection and prediction using stacked bidirectional long short term memory," *Pattern Recognition Letters*, vol. 128, pp. 529–535, 2019.
- [91] H. Khan, L. Marcuse, M. Fields, K. Swann, and B. Yener, "Focal onset seizure prediction using convolutional networks," *IEEE Transactions on Biomedical Engineering*, vol. 65, no. 9, pp. 2109–2118, 9 2018.
- [92] N. D. Truong, A. D. Nguyen, L. Kuhlmann, M. R. Bonyadi, J. Yang, and O. Kavehei, "A generalised seizure prediction with convolutional neural networks for intracranial and scalp electroencephalogram data analysis," *arXiv preprint arXiv:1707.01976*, 2017.
- [93] M. Winterhalter, T. Maiwald, H. Voss, R. Aschenbrenner-Scheibe, J. Timmer, and A. Schulze-Bonhage, "The seizure prediction characteristic: a general framework to assess and compare seizure prediction methods," *Epilepsy & Behavior*, vol. 4, no. 3, pp. 318–325, 2003.
- [94] A. H. Shueb, "Application of machine learning to epileptic seizure onset detection and treatment," Ph.D. dissertation, Massachusetts Institute of Technology, 2009.
- [95] R. Hussein, M. O. Ahmed, R. Ward, Z. J. Wang, L. Kuhlmann, and Y. Guo, "Human intracranial eeg quantitative analysis and automatic feature learning for epileptic seizure prediction," *arXiv preprint arXiv:1904.03603*, 2019.
- [96] I. Kiral-Kornek, S. Roy, E. Nurse, B. Mashford, P. Karoly, T. Carroll, D. Payne, S. Saha, S. Baldassano, T. O'Brien *et al.*, "Epileptic seizure prediction using big data and deep learning: toward a mobile system," *EBioMedicine*, vol. 27, pp. 103–111, 2018.
- [97] L. Kuhlmann, P. Karoly, D. R. Freestone, B. H. Brinkmann, A. Temko, A. Barachant, F. Li, G. Titericz Jr, B. W. Lang, D. Lavery *et al.*, "Epilepsysystem.org: crowd-sourcing reproducible seizure prediction with long-term human intracranial eeg," *Brain*, vol. 141, no. 9, pp. 2619–2630, 2018.
- [98] M. J. Cook, T. J. O'Brien, S. F. Berkovic, M. Murphy, A. Morokoff, G. Fabinyi, W. D'Souza, R. Yerra, J. Archer, L. Litewka *et al.*, "Prediction of seizure likelihood with a long-term, implanted seizure advisory system in patients with drug-resistant epilepsy: a first-in-man study," *The Lancet Neurology*, vol. 12, no. 6, pp. 563–571, 2013.
- [99] R. G. Andrzejak, K. Lehnertz, F. Mormann, C. Rieke, P. David, and C. E. Elger, "Indications of nonlinear deterministic and finite-dimensional structures in time series of brain electrical activity: Dependence on recording region and brain state," *Physical Review E*, vol. 64, no. 6, p. 061907, 2001.
- [100] N. D. Truong, L. Kuhlmann, M. R. Bonyadi, D. Querlioz, L. Zhou, and O. Kavehei, "Epileptic seizure forecasting with generative adversarial networks," *IEEE Access*, vol. 7, pp. 143 999–144 009, 2019.
- [101] M. Ihle, H. Feldwisch-Drentrup, C. A. Teixeira, A. Witon, B. Schelter, J. Timmer, and A. Schulze-Bonhage, "Epilepsiae—a european epilepsy database," *Computer methods and programs in biomedicine*, vol. 106, no. 3, pp. 127–138, 2012.



- [102] H. Gammulle, S. Denman, S. Sridharan, and C. Fookes, "Fine-grained action segmentation using the semi-supervised action gan," *Pattern Recognition*, vol. 98, p. 107039, 2020.
- [103] A. Singh, S. Sengupta, and V. Lakshminarayanan, "Explainable deep learning models in medical image analysis," *arXiv preprint arXiv:2005.13799*, 2020.
- [104] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2921–2929.
- [105] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618–626.
- [106] A. Chattopadhyay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, "Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks," in *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2018, pp. 839–847.
- [107] M. T. Ribeiro, S. Singh, and C. 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*, 2016, pp. 1135–1144.
- [108] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in neural information processing systems*, 2017, pp. 4765–4774.
- [109] S. M. Lundberg, B. Nair, M. S. Vavilala, M. Horibe, M. J. Eisses, T. Adams, D. E. Liston, D. K.-W. Low, S.-F. Newman, J. Kim *et al.*, "Explainable machine-learning predictions for the prevention of hypoxaemia during surgery," *Nature biomedical engineering*, vol. 2, no. 10, pp. 749–760, 2018.
- [110] A. Shrikumar, P. Greenside, and A. Kundaje, "Learning important features through propagating activation differences," *arXiv preprint arXiv:1704.02685*, 2017.
- [111] G. Montavon, S. Lapuschkin, A. Binder, W. Samek, and K.-R. Müller, "Explaining nonlinear classification decisions with deep taylor decomposition," *Pattern Recognition*, vol. 65, pp. 211–222, 2017.
- [112] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller, "Striving for simplicity: The all convolutional net," *arXiv preprint arXiv:1412.6806*, 2014.
- [113] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," *arXiv preprint arXiv:1703.01365*, 2017.
- [114] S. Pereira, R. Meier, V. Alves, M. Reyes, and C. A. Silva, "Automatic brain tumor grading from mri data using convolutional neural networks and quality assessment," in *Understanding and interpreting machine learning in medical image computing applications*. Springer, 2018, pp. 106–114.
- [115] K. Young, G. Booth, B. Simpson, R. Dutton, and S. Shrapnel, "Deep neural network or dermatologist?" in *Interpretability of Machine Intelligence in Medical Image Computing and Multimodal Learning for Clinical Decision Support*. Springer, 2019, pp. 48–55.
- [116] N. Tsiknakis, E. Trivizakis, E. E. Vassalou, G. Z. Papadakis, D. A. Spandidos, A. Tsatsakis, J. Sánchez-García, R. López-González, N. Papanikolaou, A. H. Karantanis *et al.*, "Interpretable artificial intelligence framework for covid-19 screening on chest x-rays," *Experimental and Therapeutic Medicine*, vol. 20, no. 2, pp. 727–735, 2020.
- [117] S. Chatterjee, F. Saad, C. Sarasaen, S. Ghosh, R. Khatun, P. Radeva, G. Rose, S. Stober, O. Speck, and A. Nürnberger, "Exploration of interpretability techniques for deep covid-19 classification using chest x-ray images," *arXiv preprint arXiv:2006.02570*, 2020.
- [118] Z. Zhang, Y. Xie, F. Xing, M. McGough, and L. Yang, "Mdnnet: A semantically and visually interpretable medical image diagnosis network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 6428–6436.
- [119] H. Lee, S. T. Kim, and Y. M. Ro, "Generation of multimodal justification using visual word constraint model for explainable computer-aided diagnosis," in *Interpretability of Machine Intelligence in Medical Image Computing and Multimodal Learning for Clinical Decision Support*. Springer, 2019, pp. 21–29.
- [120] P. Zhu and M. Ogino, "Guideline-based additive explanation for computer-aided diagnosis of lung nodules," in *Interpretability of Machine Intelligence in Medical Image Computing and Multimodal Learning for Clinical Decision Support*. Springer, 2019, pp. 39–47.
- [121] M. R. Arbabshirani, B. K. Fornwalt, G. J. Mongelluzzo, J. D. Suever, B. D. Geise, A. A. Patel, and G. J. Moore, "Advanced machine learning in action: identification of intracranial hemorrhage on computed tomography scans of the head with clinical workflow integration," *NPJ digital medicine*, vol. 1, no. 1, pp. 1–7, 2018.
- [122] A. Almazroa, S. Alodhayb, E. Osman, E. Ramadan, M. Hummadi, M. Dlam, M. Alkatee, K. Raahemifar, and V. Lakshminarayanan, "Agreement among ophthalmologists in marking the optic disc and optic cup in fundus images," *International ophthalmology*, vol. 37, no. 3, pp. 701–717, 2017.
- [123] C. Huang, Y. Wu, Y. Zuo, K. Pei, and G. Min, "Towards experienced anomaly detector through reinforcement learning," in *AAAI*, 2018.
- [124] M.-h. Oh and G. Iyengar, "Sequential anomaly detection using inverse reinforcement learning," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2019, pp. 1480–1490.
- [125] J. G. Richens, C. M. Lee, and S. Johri, "Improving the accuracy of medical diagnosis with causal machine learning," *Nature communications*, vol. 11, no. 1, pp. 1–9, 2020.
- [126] D. C. Castro, I. Walker, and B. Glocker, "Causality matters in medical imaging," *Nature Communications*, vol. 11, no. 1, pp. 1–10, 2020.
- [127] N. Altman and M. Krzywinski, "Association, correlation and causation," *Nature Methods*, 2015.
- [128] M. Nauta, D. Bucur, and C. Seifert, "Causal discovery with attention-based convolutional neural networks," *Machine Learning and Knowledge Extraction*, vol. 1, no. 1, pp. 312–340, 2019.
- [129] H. Zenil, N. A. Kiani, F. Marabita, Y. Deng, S. Elias, A. Schmidt, G. Ball, and J. Tegnér, "An algorithmic information calculus for causal discovery and reprogramming systems," *iScience*, vol. 19, pp. 1160–1172, 2019.
- [130] H. Zenil, N. A. Kiani, A. A. Zea, and J. Tegnér, "Causal deconvolution by algorithmic generative models," *Nature Machine Intelligence*, vol. 1, no. 1, pp. 58–66, 2019.
- [131] Y. Zhang, S. Pal, M. Coates, and D. Ustebay, "Bayesian graph convolutional neural networks for semi-supervised classification," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 5829–5836.
- [132] S. Pal, F. Regol, and M. Coates, "Bayesian graph convolutional neural networks using node copying," *arXiv preprint arXiv:1911.04965*, 2019.
- [133] A. Kendall and Y. Gal, "What uncertainties do we need in bayesian deep learning for computer vision?" in *Advances in neural information processing systems*, 2017, pp. 5574–5584.
- [134] C. Leibig, V. Allken, M. S. Ayhan, P. Berens, and S. Wahl, "Leveraging uncertainty information from deep neural networks for disease detection," *Scientific reports*, vol. 7, no. 1, pp. 1–14, 2017.
- [135] Y. Kwon, J.-H. Won, B. J. Kim, and M. C. Paik, "Uncertainty quantification using bayesian neural networks in classification: Application to biomedical image segmentation," *Computational Statistics & Data Analysis*, vol. 142, p. 106816, 2020.
- [136] T. DeVries and G. W. Taylor, "Leveraging uncertainty estimates for predicting segmentation quality," *arXiv preprint arXiv:1807.00502*, 2018.
- [137] P. Seeböck, J. I. Orlando, T. Schlegl, S. M. Waldstein, H. Bogunović, S. Klimesch, G. Langs, and U. Schmidt-Erfurth, "Exploiting epistemic uncertainty of anatomy segmentation for anomaly detection in retinal oct," *IEEE transactions on medical imaging*, vol. 39, no. 1, pp. 87–98, 2019.
- [138] T. Nair, D. Precup, D. L. Arnold, and T. Arbel, "Exploring uncertainty measures in deep networks for multiple sclerosis lesion detection and segmentation," *Medical image analysis*, vol. 59, p. 101557, 2020.
- [139] T. Dissanayake, T. Fernando, S. Denman, H. Ghaemmaghami, S. Sridharan, and C. Fookes, "Domain generalization in biosignal classification," *arXiv preprint arXiv:2011.06207*, 2020.
- [140] S. Hu, J. Tomczak, and M. Welling, "Meta-learning for medical image classification," *Conference on Medical Imaging with Deep Learning (MIDL 2018)*, 2018.
- [141] K. Mahajan, M. Sharma, and L. Vig, "Meta-dermdagnosis: Few-shot skin disease identification using meta-learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020, pp. 730–731.
- [142] X. S. Zhang, F. Tang, H. H. Dodge, J. Zhou, and F. Wang, "Metapred: Meta-learning for clinical risk prediction with limited patient electronic health records," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2019, pp. 2487–2495.
- [143] J. M. Johnson and T. M. Khoshgoufar, "Survey on deep learning with class imbalance," *Journal of Big Data*, vol. 6, no. 1, p. 27, 2019.
- [144] L. Zhang, H. Yang, and Z. Jiang, "Imbalanced biomedical data classification using self-adaptive multilayer elm combined with dynamic gan," *Biomedical engineering online*, vol. 17, no. 1, p. 181, 2018.

- [145] S. S. Mullick, S. Datta, and S. Das, “Generative adversarial minority oversampling,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 1695–1704.
- [146] T. Zhou, W. Liu, C. Zhou, and L. Chen, “Gan-based semi-supervised for imbalanced data classification,” in *2018 4th International Conference on Information Management (ICIM)*. IEEE, 2018, pp. 17–21.
- [147] C. Doersch, A. Gupta, and A. A. Efros, “Unsupervised visual representation learning by context prediction,” in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1422–1430.
- [148] R. Zhang, P. Isola, and A. A. Efros, “Colorful image colorization,” in *European conference on computer vision*. Springer, 2016, pp. 649–666.
- [149] M. Noroozi and P. Favaro, “Unsupervised learning of visual representations by solving jigsaw puzzles,” in *European Conference on Computer Vision*. Springer, 2016, pp. 69–84.
- [150] R. Ali, M. U. K. Khan, and C. M. Kyung, “Self-supervised representation learning for visual anomaly detection,” *arXiv preprint arXiv:2006.09654*, 2020.