Medtronic
Engineering the extraordinary

SAIL Lab

Comparative Study of Generative Models for Early Detection of Failures in Medical Devices

First International Medical Device Safety Risk Management Conference

April 2024

Binesh Kumar*, Sr. Principal R&D Engineer, Medtronic
Bahareh Arghavani Nobar, Graduate Research Assistant, SAIL Lab
Dr. Vahid Behzadan, Director, SAIL Lab, University of New Haven

Agenda

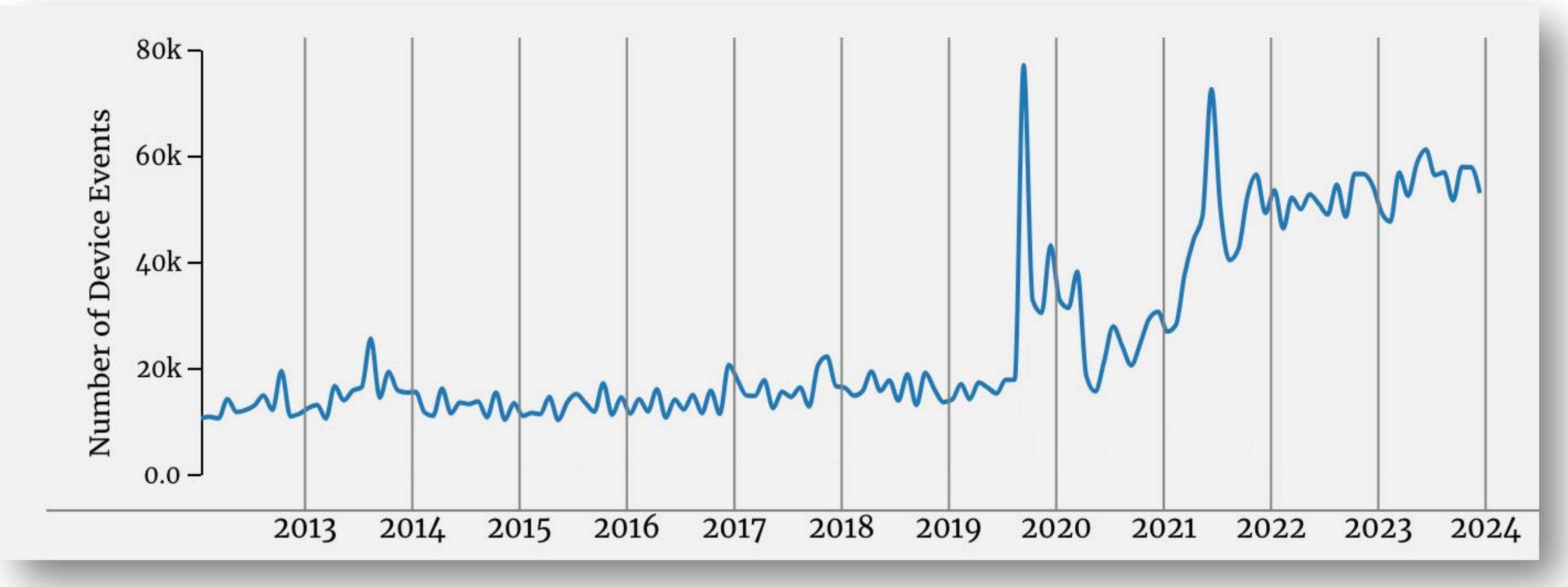
1. Introduction
2. Research Problem
3. Research Methodology
4. Review of the state of the art
5. Study Setup
6. Study Results
7. Conclusion & Next Steps

Introduction

The Problem

SAIL Lab

Fig:1 Device Failures Resulting in an Injury



Source: FDA Open Access - MAUDE

- ➔ 2006-2011: 5,294 device recalls, 1.15M adverse events, 92,600 injuries, 25,800 deaths [Ale+13]
- ➔ 2017: 627 software devices recalled, 12 high-risk [RZ17]
- ➔ 2018-2020: 383 stapler complications, 22 deaths [CB22]

Medtronic

Introduction

The Problem

- ➔ Fault diagnosis is a seasoned field of research, and many critical medical devices maintain an embedded expert system to self diagnose
- ➔ Increasing complexity of the devices and the variability in the operating environment results in failures which are often hard to predict and prevent by expert systems
- ➔ Majority of medical devices embed intricate electro-mechanical components with varying properties where failure is not linear
- ➔ The increasing complexity of medical devices, particularly in their electronics and software components, presents significant challenges in terms of safety, reliability, and efficacy [JSR13]self-diagnose

Definitions

- *Fault: An unpermitted deviation of at least one characteristic property or parameter from the expected normal condition in a piece of equipment or a system that may lead to a failure*
- *Fault detection: Fault detection is monitoring approaches to detect, isolate, and identify the faults by using the concept of redundancy, either hardware redundancy or analytical redundancy*
- *Failure Prediction: Prediction of the future status of the faulty component(s) and estimation of the remaining useful lifetime based on the available information. The prognosis task allows predicting the future state of damage, rather than diagnosing the current state of damage*
- *Anomaly detection: refers to the problem of finding patterns in data that do not conform to expected behavior*

Introduction

Study Framework

Research Problem

Given the increasing complexity of medical devices and the rising incidence of device-related failures and recalls, there is a pressing need to develop and implement advanced generative methods for more effective and reliable fault detection in these devices.

Literature Review

Conduct a literature review on medical device failures and fault detection techniques, focusing on the role of generative models in prediction and prevention.

Framework Development

Establish the theoretical foundation, research objectives and hypotheses.

Analyze two generative algorithms (GAN, VAE) and one classic (HMM) for their effectiveness in fault detection.

Preliminary Study

Evaluated the applicability against

- ○ Real world Surgical device sensor data
- ○ Airbus anomaly detection as benchmark dataset

Introduction

Team

SAIL Lab

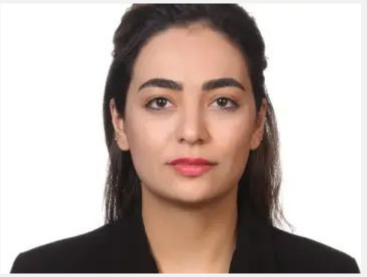
- Quality is a cornerstone of the Surgical Stapling mission, and we believe that with predictive and preventive maintenance we would further enhance our customer experience
- Stapling Research & Technology team partnered with University of New Haven's Secured and Assured Intelligence Lab (SAIL) for an independent study to survey the state-of-the-art Fault Detection and Prediction algorithms and perform a feasibility study to understand the applicability to medical devices.

Binesh Kumar Sr Prin. Engineer, ASI	Principal Investigator
Andrew Miesse Director, R&T Surgical	Advisor & Sponsor

Dr. Vahid Behzadan — Principal Investigator

Vahid Behzadan is an Assistant Professor in Computer Science and Data Science. He is also the founder and director of the Secure and Assured Intelligent Learning (SAIL) research group, and mentors the University's hacking team. Dr. Behzadan's research is primarily on the safety and security of artificial intelligence and complex adaptive systems. His pioneering work on the security of deep reinforcement learning is recognized as seminal contributions to this growing field of research

Bahareh Arghavani Nobar — Research Assistant



She is currently a Master's in Data Science student at the University of New Haven and works as a research assistant at SAIL Lab, where her primary focus is on AI applications in healthcare.

Introduction

Literature Review

SAIL Lab

Literature Survey

- Semi-systematic literature survey for each of the below topics, qualitative/quantitative analysis, the development over time, and synthesizing and comparing evidence using data generated by a medical device
  - Survey of fault detection approaches in medical devices, established benchmarks, datasets, and tools
  - Survey of supervised and unsupervised machine learning techniques
  - Survey of deep generative and Reinforcement Learning(RL) based approaches to fault detection and evaluation of an online RL + Generative Adversarial Networks approach to fault prediction
  - Survey of Koopman operators for fault detection as well as a novel formulation of Koopman approach to fault prediction in medical devices

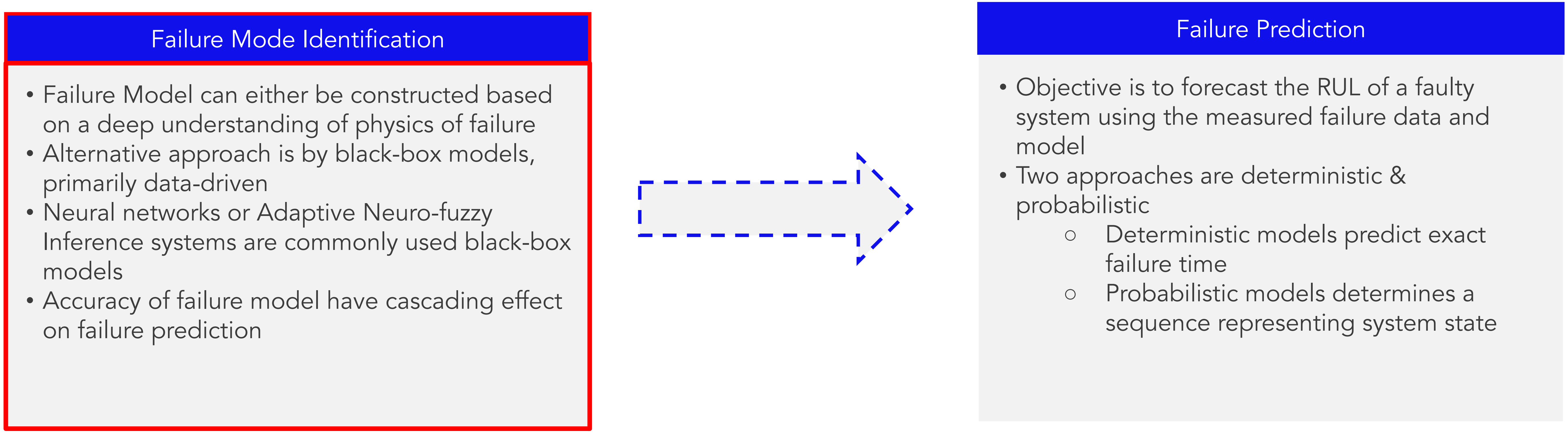
Databases

- IEEE, Scopus databases and proceedings from International Symposium on Reliable Distributed Systems, International Conference on Software Engineering, International Conference on Machine Learning and Application, Conference on Neural Information Processing Systems and Association for the Advancement of Artificial Intelligence conferences

Literature Review

Fault Diagnosis & Prognosis

- ➔ An effective Fault Diagnosis & Prognosis system requires fast and reliable fault detection, isolation and prognosis
- ➔ Fault prognosis is challenging due to uncertainties in the system, prediction horizon of the failure and lack of measurements directly related to the dynamics of failures



Failure Mode Identification

- Failure Model can either be constructed based on a deep understanding of physics of failure
- Alternative approach is by black-box models, primarily data-driven
- Neural networks or Adaptive Neuro-fuzzy Inference systems are commonly used black-box models
- Accuracy of failure model have cascading effect on failure prediction

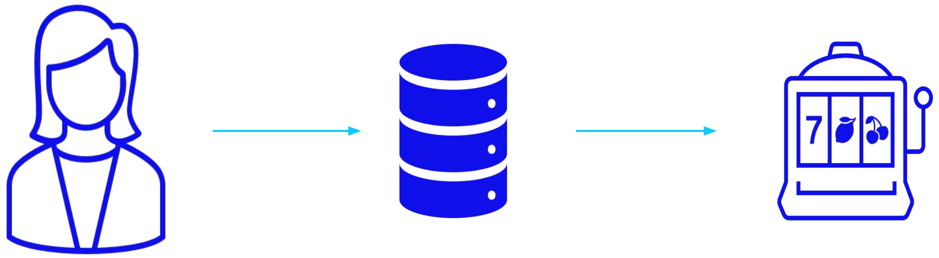
Failure Prediction

- Objective is to forecast the RUL of a faulty system using the measured failure data and model
- Two approaches are deterministic & probabilistic
  - ○ Deterministic models predict exact failure time
  - ○ Probabilistic models determines a sequence representing system state

Literature Review

Fault Mode Identification - Knowledge Based Systems

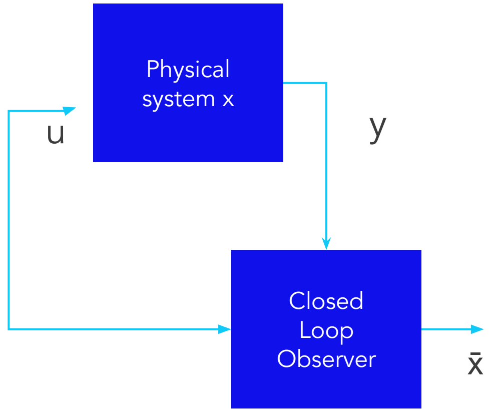
- ➔ Medical devices widely use Knowledge-Based systems for fault detection and as decision support tools [MartinT91]
- ➔ These systems include a knowledge database and an inference engine to predict anomalies, indicating potential faults.
- ➔ Their popularity stems from reliability and simplicity, but creating a complete rule set for complex systems can be challenging.
- ➔ Traditional model-based techniques are also common for product reliability testing during the development phase [MartinT91]

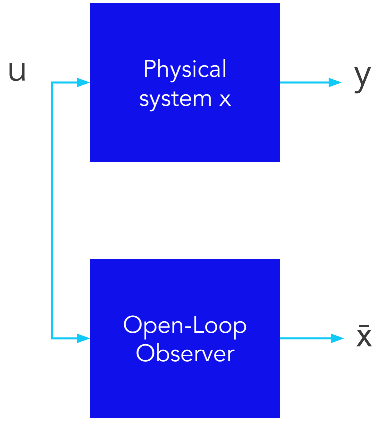


Literature Review

Fault Mode Identification - Observer Methods

- ➔ Sensor-reliant critical devices often use observer-based methods, calculating residuals from system outputs and observer estimates.
- ➔ Common approaches include dedicated observer schemes and robust sliding mode observers, both closed loop and multiple open-loop types.
- ➔ There's a balance between estimation accuracy and fault detection capability in these methods.

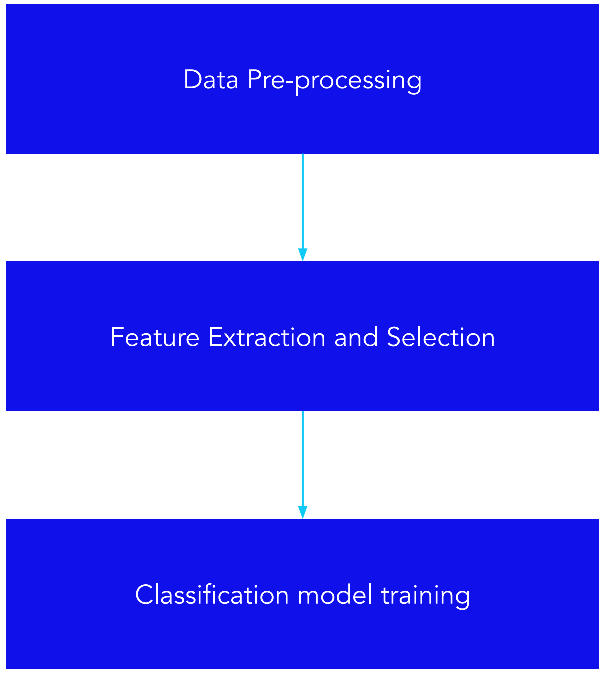




Literature Review

Supervised Fault Detection

- ➔ Fault detection using supervised learning technique involves feature extraction, selection and fault classification
- ➔ Feature extraction improves the information density – kernel based PCA is very commonly used in fault detection, FFT (Fast Fourier Transform) is often used for time series [DSJN19]
- ➔ Feature selection methods aims to improve the quality of data, supervised methods such as Correlation Based Feature Selection and Fast Correlation Filter are widely used [YuL03].
- ➔ Popular feature selection methods for time series are mean, variance, kurtosis and for sequences distance-based entropy is employed [TsimpirisA12]



Data Pre-processing

Feature Extraction and Selection

Classification model training

Literature Review

Supervised Fault Detection

SAIL Lab

- ➔ Popular supervised learning techniques for fault detection are linear and logistic regression models, support vector machines, decision tree, clustering techniques and Naïve Bayes [HaTMP19]
- ➔ SVM is a popular choice among researchers for time series classification [HaTMP19]
- ➔ Multi-class approach using SVM is computationally expensive and is often challenging for online classification based on a study by Cheng Jing et. Al
- ➔ In real-world there is scarce of of labeled anomalous training samples often leads to class imbalance

Medtronic

Research Methodology

Stapler Dataset

- ➔ The Surgical Stapler Dataset includes time series signals from a motorized laparoscopic stapler, with 3000 samples per instance.
- ➔ Captures three distinct features at a 1 kHz sampling rate in benchtop settings.
- ➔ Uses high-speed camera data to accurately mark anomalies and faults for precise fault detection analysis.

Fig:2 Staple Motor Current Data - Normal

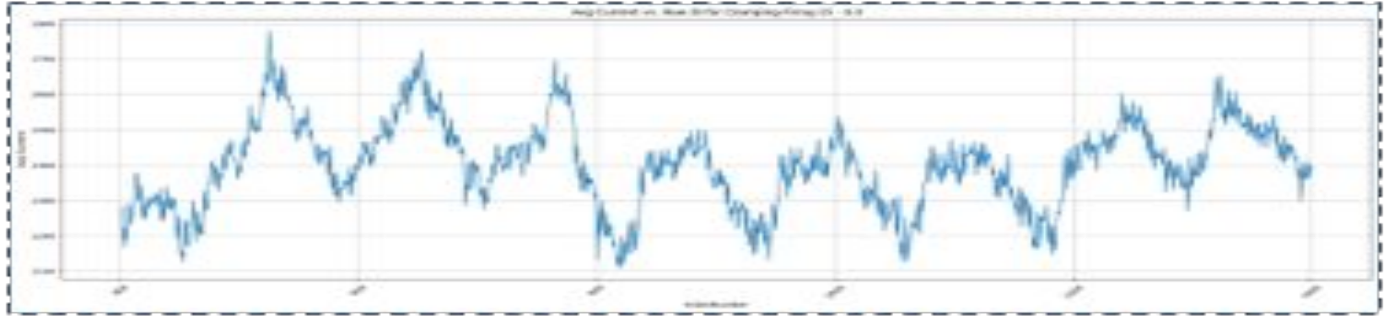
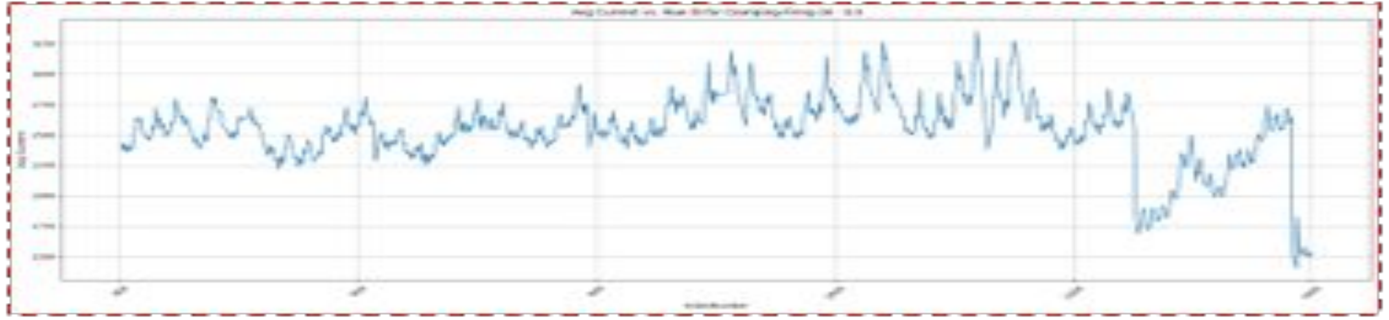


Fig:3 Staple Motor Current Data - Faulty

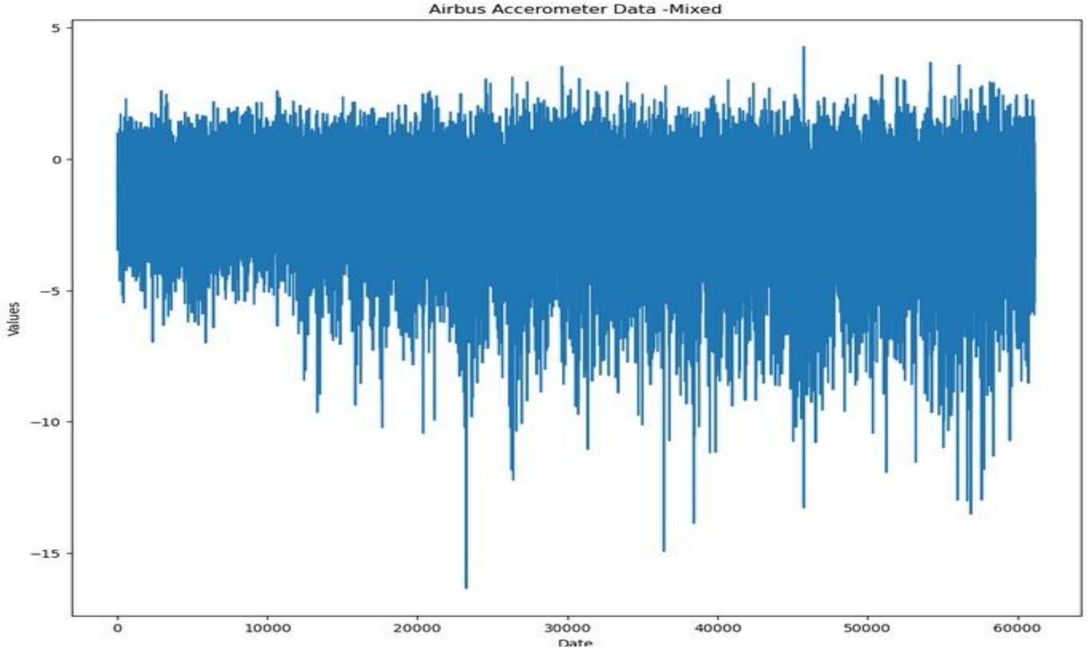


Research Methodology

Airbus Dataset

- ➔ The dataset serves as a baseline for fault detection, capturing vibrations from real helicopter flight tests using accelerometers placed at various positions.
- ➔ Includes 1,677 sequences of normal flights for training, which are essential for understanding standard vibration patterns.
- ➔ For testing anomaly detection, it provides a balanced test set with 594 sequences, encompassing both normal and abnormal flight conditions.

Fig:4 Airbus Helicopter Vibration Data - Mix

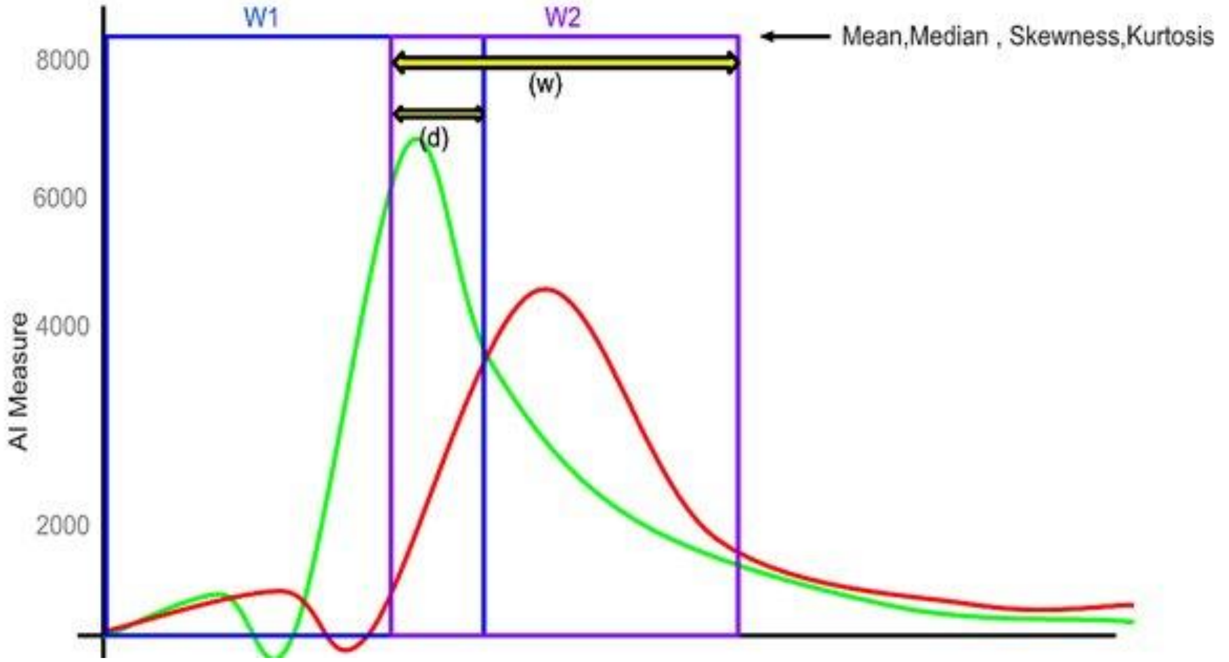


Research Methodology

Data Pre-Processing

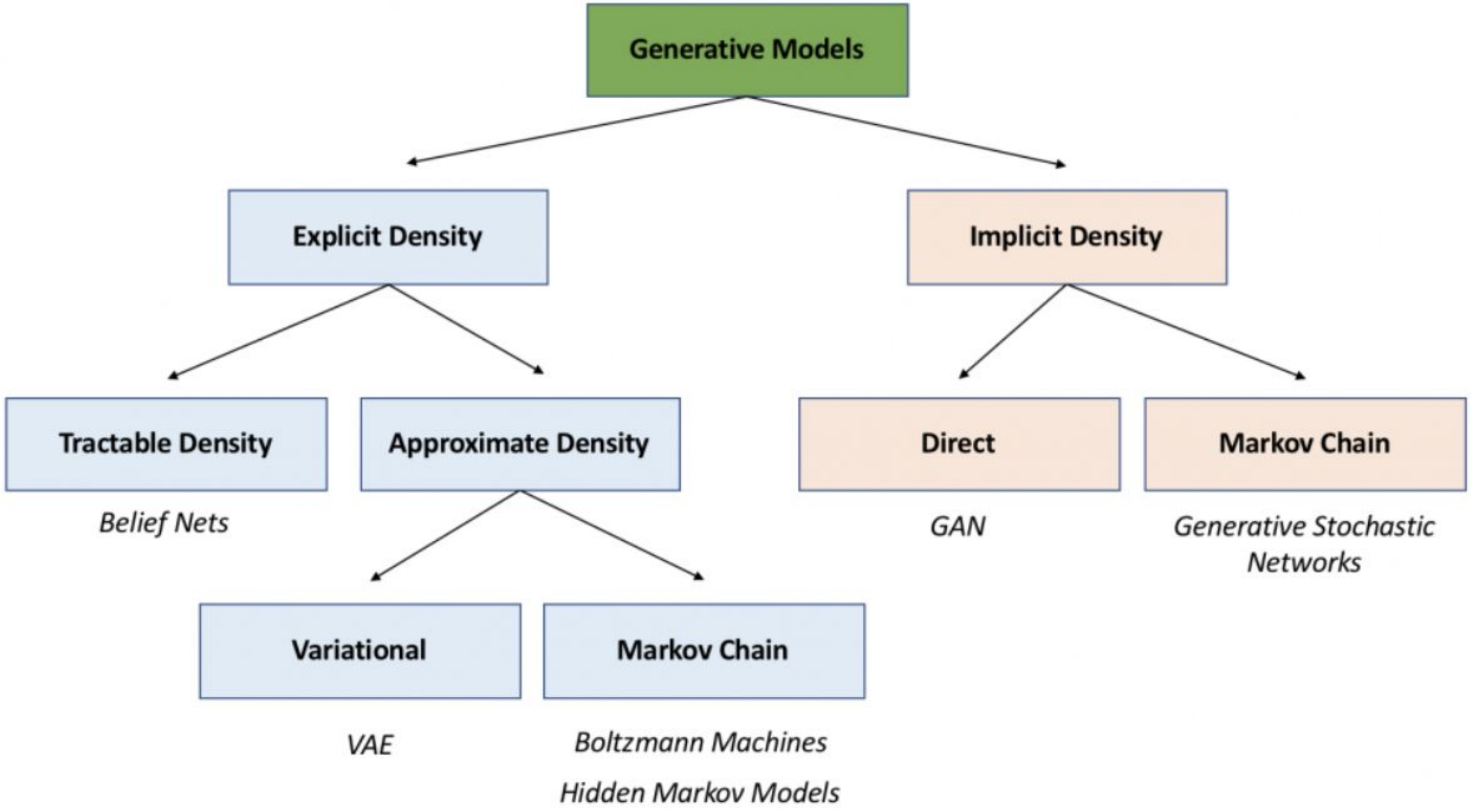
- ➔ Sequence Padding: Standardized sequence lengths with padding, creating a uniform input for model training.
- ➔ Windowing: Segmented continuous data into fixed-size windows for localized analysis.
- ➔ Feature Extraction: Computed statistical features (mean, median, skewness, kurtosis) within each window to capture data characteristics.
- ➔ Data Splitting: Divided data into training (90% normal samples) and testing sets (remaining normal and all faulty samples).
- ➔ Data Scaling: Implemented standardization using MinMax for stable and efficient model training.

Fig:6 Windowing Approach In Pre-Processing



Deep Generative Model Approaches

SAIL Lab



Deep Generative Model Approaches

Unsupervised Learning

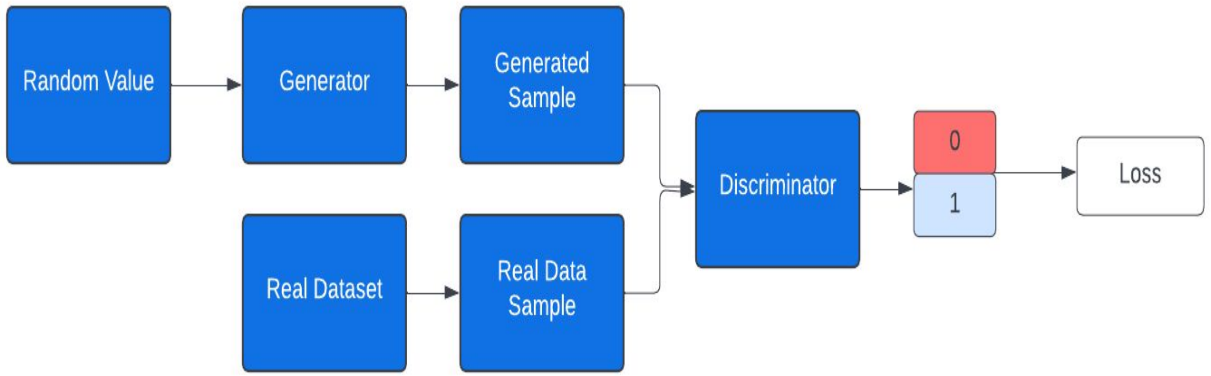
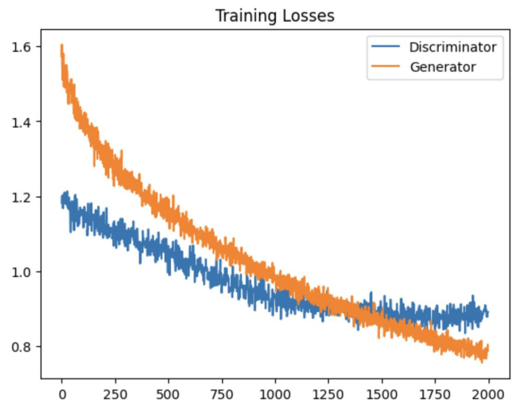


Fig:7.a GAN Architecture



$$\min_G \max_D V(D,G) = \mathbb{E}_{\mathbf{x}\sim p_{data}(\mathbf{x})}[\log D(\mathbf{x})] + \mathbb{E}_{\mathbf{z}\sim p_z(\mathbf{z})}[\log(1 - D(G(\mathbf{z})))] .$$

Fig:7.b GAN Architecture

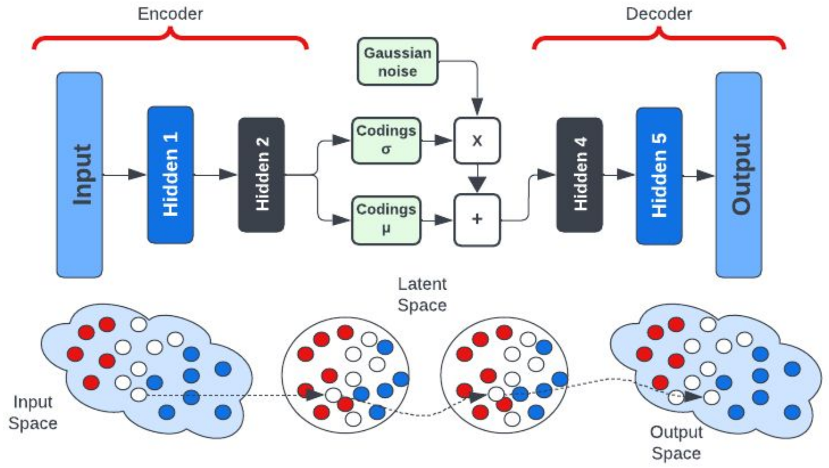
- ➔ GANs operate on an adversarial process involving two models: a discriminator (D) and a generator (G), trained simultaneously in a zero-sum game.
- ➔ The generator's goal is to capture the data distribution, while the discriminator evaluates whether a sample is from the training data or the generator.
- ➔ Training GANs only on normal samples enables the generator to learn normal data patterns, leading to increased reconstruction error for anomalous samples
- ➔ The best threshold is selected by sweeping through a range of possible thresholds and computing various evaluation metrics (such as accuracy, precision, recall, and F1-score) for each threshold. The threshold that yields the highest F1-score is chosen as the optimal threshold.

Deep Generative Model Approaches

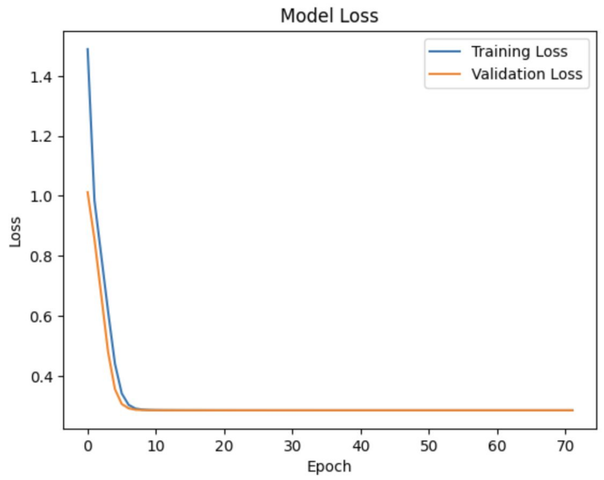
Unsupervised Learning

- ➔ VAEs are likelihood-based generative models that maximize the training data likelihood using the model pθ(Data).
- ➔ They use latent variables z to compute pθ(x), integrating over pθ(x|z)p(z)dz.
- ➔ VAEs are trained efficiently via an inference distribution qφ(z|x).
- ➔ Anomalies in VAEs are identified by high reconstruction error or low probability, as they represent deviations from the learned distribution.
- ➔ The best threshold for anomaly detection is set by evaluating different thresholds and choosing the one that maximizes the F1 score which is (6.92), balancing precision and recall.

Fig:8 VAE Architecture



$$L_{\mathrm{VAE}} = \mathbb{E}_{q(z|x)}\left[\log p(x|z)\right] - D_{\mathrm{KL}}\left(q(z|x)||p(z)\right)$$

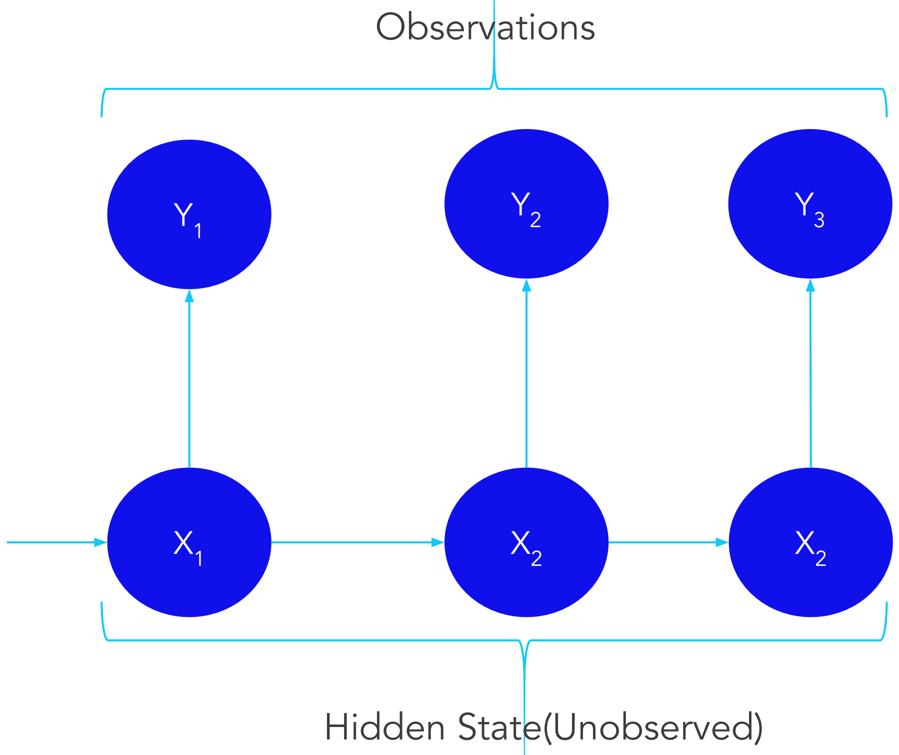


Classic Generative Model Approaches

Unsupervised Learning

- ➔ HMMs are preferred for statistically modeling linear time series or sequence problems.
- ➔ They consist of an unobservable stochastic process, inferred through a sequence of observable symbols.
- ➔ Defined by λ = (A, B, π), where A is the state transition probability matrix, B the observation symbol probability matrix, and π the initial state probability vector.
- ➔ HMM-based anomaly detection involves training on normal observations, calculating normal behavior states, and comparing these to states from new data for anomaly identification.
- ➔ The best threshold is found by trying different percentiles of log-likelihood values and evaluating which one gives the highest F1 score which is 21 percentile, ensuring a good balance between detecting anomalies (recall) and avoiding false alarms (precision)

Fig:9 HMM Architecture



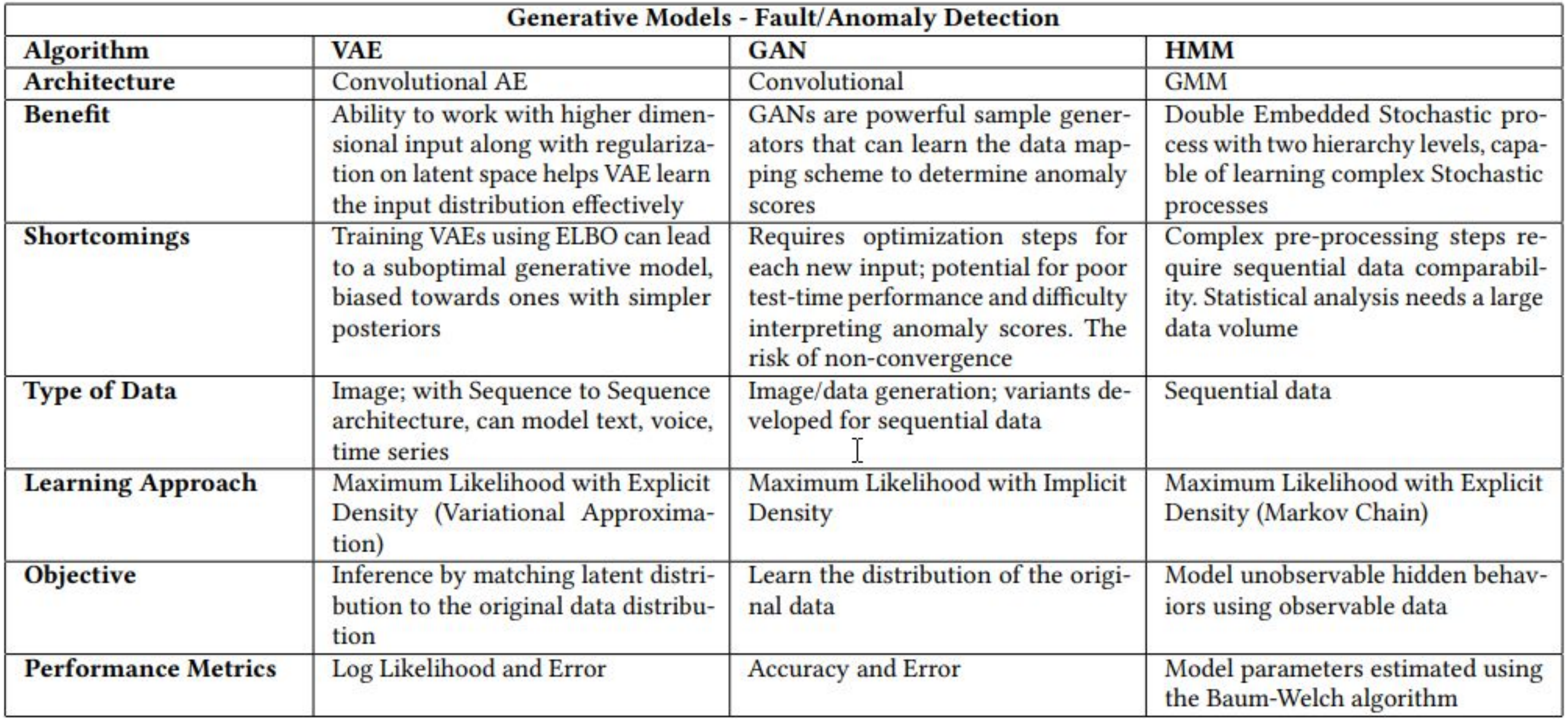
$$P(X_{1:T}, Y_{1:T}) = P(X_1)P(Y_1|X_1) \prod^{T}_{t=2} P(X_t|X_{t-1})P(Y_t|X_t)$$

Generative Model Approaches

Comparison

SAIL Lab

Fig:10 Generative Model Approaches - Comparison



Generative Models - Fault/Anomaly Detection

Algorithm	**VAE**	**GAN**	**HMM**
Architecture	Convolutional AE	Convolutional	GMM
Benefit	Ability to work with higher dimensional input along with regularization on latent space helps VAE learn the input distribution effectively	GANs are powerful sample generators that can learn the data mapping scheme to determine anomaly scores	Double Embedded Stochastic process with two hierarchy levels, capable of learning complex Stochastic processes
Shortcomings	Training VAEs using ELBO can lead to a suboptimal generative model, biased towards ones with simpler posteriors	Requires optimization steps for each new input; potential for poor test-time performance and difficulty interpreting anomaly scores. The risk of non-convergence	Complex pre-processing steps require sequential data comparability. Statistical analysis needs a large data volume
Type of Data	Image; with Sequence to Sequence architecture, can model text, voice, time series	Image/data generation; variants developed for sequential data	Sequential data
Learning Approach	Maximum Likelihood with Explicit Density (Variational Approximation)	Maximum Likelihood with Implicit Density	Maximum Likelihood with Explicit Density (Markov Chain)
Objective	Inference by matching latent distribution to the original data distribution	Learn the distribution of the original data	Model unobservable hidden behaviors using observable data
Performance Metrics	Log Likelihood and Error	Accuracy and Error	Model parameters estimated using the Baum-Welch algorithm

Medtronic

Deep Generative Model Approaches

Model Training

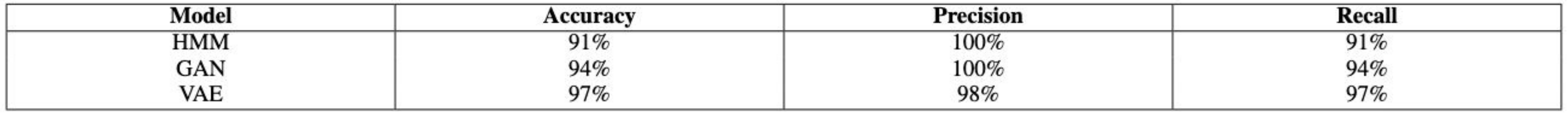
SAIL Lab

- ➔ HMM:
  - ◆ Utilized a binary state model with 'normal' and 'anomalous' states for simplicity and real-time efficiency.
  - ◆ We experimented with different thresholds by testing various percentiles of log-likelihood values. The best threshold was found at the 21st percentile of log-likelihood, which yielded the highest F1 score, effectively balancing the detection of anomalies (recall) while minimizing false alarms (precision). This approach ensured a robust and efficient anomaly detection system, tuned to perform well in real-time scenarios.
- ➔ VAE:
  - ◆ Architecture includes an encoder and decoder with ReLU and sigmoid activations and Dropout to prevent overfitting.
  - ◆ Optimized using a composite loss function (reconstruction loss + KL divergence) for enhanced data reconstruction.
- ➔ GAN:
  - ◆ Consists of a generator and discriminator with layered architectures using Relue and Tanh activations in the generator and LeakyReLU activations in the discriminator, along with sigmoid in the final layer for binary classification. BatchNormolizer in generator and Dropout is used for regularization in the discriminator.
  - ◆ Both are optimized using the Adam optimizer, and the loss function used is binary_crossentropy.

Deep Generative Model Approaches

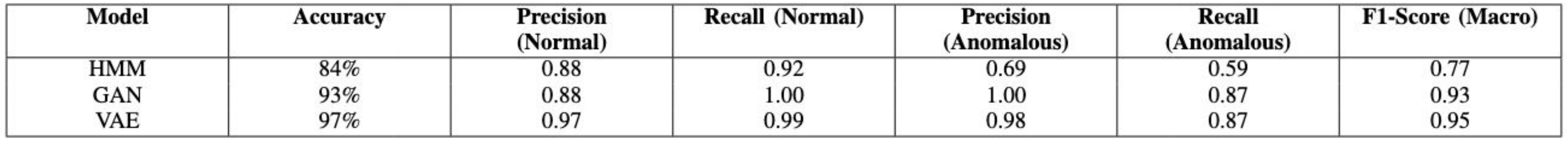
Model Performance

TABLE II: Anomaly Detection Performance on Airbus Dataset



Model	Accuracy	Precision	Recall
HMM	91%	100%	91%
GAN	94%	100%	94%
VAE	97%	98%	97%

TABLE III: Anomaly Detection Performance on Stapler Dataset



Model	Accuracy	Precision (Normal)	Recall (Normal)	Precision (Anomalous)	Recall (Anomalous)	F1-Score (Macro)
HMM	84%	0.88	0.92	0.69	0.59	0.77
GAN	93%	0.88	1.00	1.00	0.87	0.93
VAE	97%	0.97	0.99	0.98	0.87	0.95

- ➔ Benchmarking Metrics: Accuracy measures the overall correctness, precision assesses the ability to avoid false positives, and recall indicates effectiveness in identifying true positives.
- ➔ Performance Overview: Comparative results are presented in above table, showcasing each model's strengths and weaknesses in fault detection, based on the Airbus and Stapler datasets
- ➔ Generative Methods' Superiority: The research highlighted the exceptional performance of generative models, especially GAN and VAE, with a window size of 1024 ms (HMM and GAN) and for VAE we utilized 2048 ms window size, showcasing their potential in real-world anomaly detection scenarios
- ➔ Model-Specific Outcomes: The VAE model demonstrated a notable accuracy of 97%, while GAN achieved 93% accuracy. The HMM model also performed well, with an accuracy rate of 84%, underlining the effectiveness of these models compared to previous studies

Deep Generative Model Approaches

Conclusion and Future Work

SAIL Lab

➔ Study Conclusion:

- Real world comparison of GANs, VAEs, and HMMs for fault detection in Medical Device
- The superior performance of deep generative models (like GANs and VAEs) compared to HMMs on the stapler dataset suggests the need for complex models to accurately capture the stapler's intricate dynamics.
- High accuracy in learning and detecting anomalies in time-series data

➔ Future Research Directions:

- Aim for broader practical implementation in real-world medical settings
- Define the cause of anomalies by using causal models to identify which part of the device is responsible for the fault.

➔ Challenges with Generative Models:

- Computationally intensive, requiring significant hardware resources
- Large data requirements may pose challenges due to medical data privacy
- Risk of overfitting, crucial to address for medical application accuracy

Deep Generative Model Approaches

Data Availability & Acknowledgement

SAIL Lab

- We would like to extend our heartfelt gratitude to Medtronic's Surgical Operating Unit for granting us access to the data essential for this experiment. Their support has been invaluable in the advancement of this research.
- It is important to note that the data obtained from the medical device were sourced from a lab experiment and are not clinical data. Furthermore, all real values in the experiment have been anonymized for data protection and privacy
- Resources & Data
  - Airbus Helicopter Accelerometer Dataset: This dataset is publicly available and can be accessed from ETH Zürich's Research Collection - https://www.research-collection.ethz.ch/handle/20.500.11850/415151
  - Surgical Stapler Dataset: Due to security and privacy considerations, the Surgical Stapler dataset is not publically available
  - The code developed for this study is available in a GitHub repository and can be accessed via the following link https://github.com/barghavanii/Anomaly-detection-in-medical-devices-with-generative-models-/tree/main

Medtronic

THANK YOU

For any questions, please reach out to Bahareh.arghavani@gmail.com
Link to Project Site -
https://sail-lab.org/portfolio/fault-detection-and-prognosis-in-medical-devices/
Portfolio: https://barghavanii.github.io/

Fault Detection and Prediction

References

- [WZ89] K Wawryn and W Zinka. \A prototype expert system for fault diagnosis in electronic devices". In: 1989 European Conference on Circuit Theory and Design. IET. 1989, pp. 677{680
- [Fra96] Paul Martin Frank. \Analytical and qualitative model-based fault diagnosis{a survey and some new results". In: European Journal of control 2.1 (1996), pp. 6{28.
- [HR10] Vaishali Hegde and Dev Raheja. \Design for reliability in medical devices". In: 2010 Proceedings-Annual Reliability and Maintainability Symposium (RAMS). IEEE. 2010, pp. 1{6.
- [Abu17] Bassem Abu-Nasser. \Medical expert systems survey". In: International Journal of Engineering and Information Systems (IJEAIS) 1.7 (2017), pp. 218{224.
- [Kor+19] Mojtaba Kordestani et al. \Failure prognosis and applications|A survey of recent literature". In: IEEE transactions on reliability (2019).
- [Son+20] Wenyan Song et al. \Human factors risk assessment: An integrated method for improving safety in clinical use of medical devices". In: Applied Soft Computing 86 (2020), p. 105918.
- [GG] Aditya Gulati and Jeetsagar Ghorai. \Prognostic Health Management for Turbofan Engines"
- [CB22] Benjamin Clapp, et al. "Stapler malfunctions in bariatric surgery: an analysis of the MAUDE database." In: JSLS: Journal of the Society of Laparoscopic & Robotic Surgeons 26.1 (2022)
- [JSR13] Raoul Jetley, Sithu Sudarsan, and Srini Ramaswamy. "Medical Software–Issues and Best Practices." In: Proceedings of the 9th International Conference on Distributed Computing and Internet Technology (ICDCIT 2013), Bhubaneswar, India, February 5-8, 2013. Springer Berlin Heidelberg, 2013.
- [KM19] Mojtaba Kordestani, et al. "Failure prognosis and applications—A survey of recent literature." IEEE Transactions on Reliability, vol. 70, no. 2, 2019, pp. 728-748.
- [AH13] Homa Alemzadeh, et al. "Analysis of safety-critical computer failures in medical devices." IEEE Security & Privacy, vol. 11, no. 4, 2013, pp. 14-26

Fault Detection and Prediction

References

- [DSJN19] Matheus Maia de Souza, João Cesar Netto, and Renata Galante. "FFT-2PCA: A New Feature Extraction Method for Data-Based Fault Detection." In: Database and Expert Systems Applications: 30th International Conference, DEXA 2019, Linz, Austria, August 26–29, 2019, Proceedings, Part I, vol. 30. Springer International Publishing, 2019.
- [YuL03] Lei Yu, and Huan Liu. "Feature selection for high-dimensional data: A fast correlation-based filter solution." In: Proceedings of the 20th International Conference on Machine Learning (ICML-03), 2003
- [TsimpirisA12] Alkiviadis Tsimpiris, and Dimitris Kugiumtzis. "Feature selection for classification of oscillating time series." Expert Systems, vol. 29, no. 5, 2012, pp. 456-477.
- [MartinT91] T. Patrick Martin, et al. "A knowledge-based system for fault diagnosis in real-time engineering applications." In: Database and Expert Systems Applications: Proceedings of the International Conference in Berlin, Federal Republic of Germany, 1991. Springer Vienna, 1991.
- [HaTMP19] Thi Minh Phuong Ha, et al. "Experimental study on software fault prediction using machine learning model." In: 2019 11th International Conference on Knowledge and Systems Engineering (KSE). IEEE, 2019.
- [BooyseWH20] Wihan Booyse, et al. "Deep digital twins for detection, diagnostics and prognostics." In: Mechanical Systems and Signal Processing 140 (2020): 106612