ADVANCING LOG ANOMALY DETECTION USING DEEP LEARNING AND TRANSFORMER BASED AI MODELS

MASTER OF SCIENCE ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING

LIVERPOOL JOHN MOORES UNIVERSITY

SHASHANK BHATNAGAR

STUDENT ID - 1080414

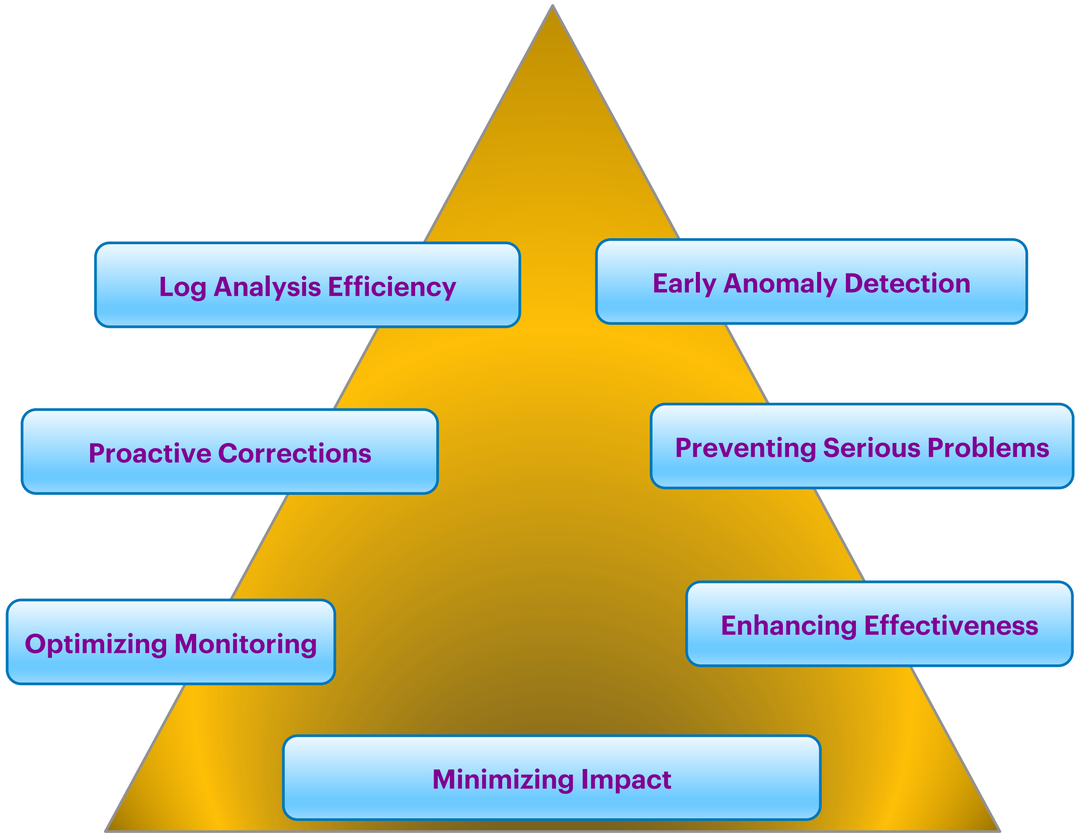
AUGUST 2023

AGENDA

- Motivation
- Background
- Problem statement
- Aim and Objectives
- Research Questions
- Literature Review
- Research Methodology
- Exploratory Data Analysis
- Feature Engineering (Log Parsing using Drain and Sliding Window)
- Model Experimentations and Results
- Model Evaluation and Summary
- Conclusions and Future Recommendations
- Limitations

MOTIVATION AND BACKGROUND

MOTIVATION



BACKGROUND

Importance of Anomaly Detection

- Anomaly detection is essential in cybersecurity and system monitoring.
- It identifies unusual data patterns, indicating potential security threats, system failures, or abnormal events.
- Benefits include threat detection, system reliability, operational efficiency, cost reduction, and regulatory compliance.

Challenges with Traditional Methods

- Traditional methods can be inefficient, time-consuming, and struggle with unstructured log data.
- Challenges include robustness, false positives, and real-time processing.
- Machine learning-based methods have emerged but may misclassify normal noise logs.

Proposed Log Sequence LSTM Approach

- A log sequence LSTM-based method is proposed to address challenges.
- It considers robustness, concept drift, and noise issues.
- Aims to automate anomaly detection in system log files, reducing manual efforts.

Utilization of Transformer-Based Methods

- A transformer-based methods for real-time log processing.
- Aims to classify anomalies, enhance system reliability, and security.

Problem Statement

Gap in Robustness and Accuracy:

- Existing anomaly detection methods face challenges in robustness and accuracy.
- Specifically, when applied to unstructured system logs, these methods encounter limitations.

Challenges in Log Data:

- Unstructured System Logs: Log data lacks structure, making analysis complex.
- High Volume of Logs: Large volumes of logs are generated daily, overwhelming manual analysis.
- Concept Drift: Changing log patterns and system behaviours pose challenges to detection methods.
- Noise Problems: Unwanted or irrelevant log entries introduce noise, affecting accuracy.

Linkage Between Anomaly Detection and ML:

- The existing literature discusses anomaly detection and machine learning separately.
- There's often a lack of direct integration, hindering the development of effective anomaly detection models..

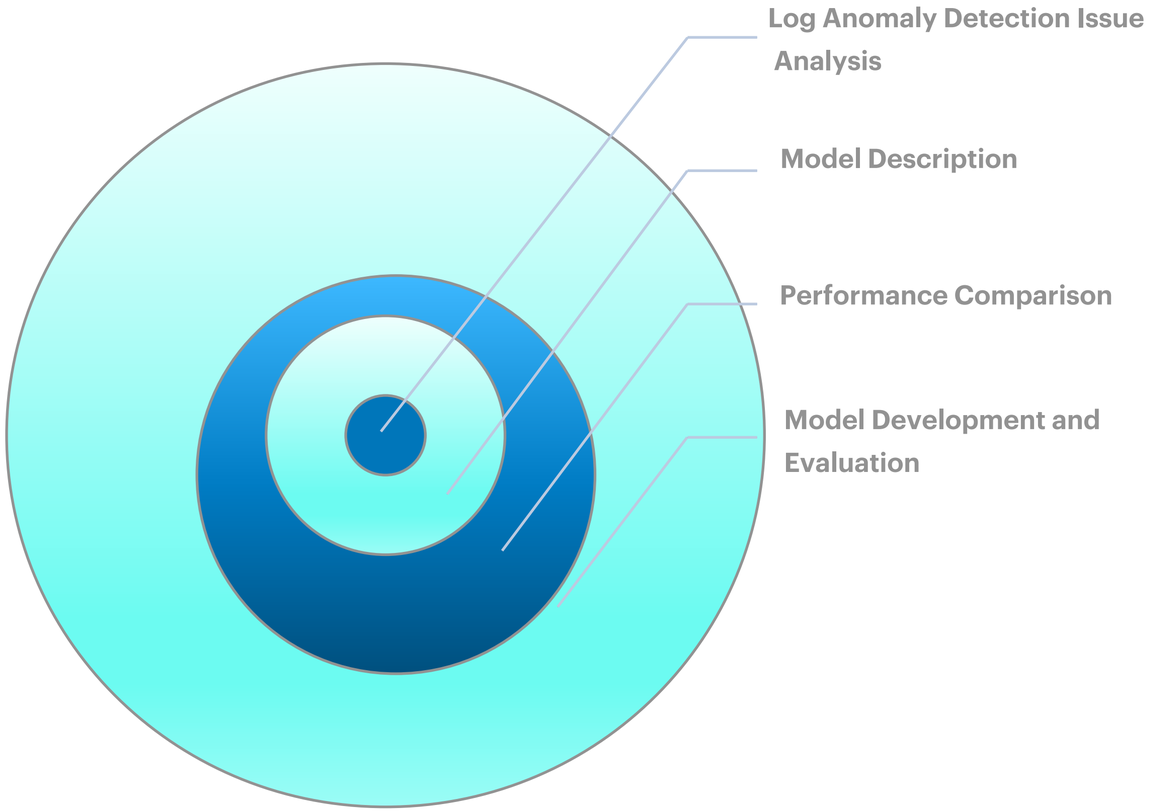
Aim & Objectives

AIM

Aim : To propose a robust model for detecting anomalies in system and event logs, enabling the identification of system malfunctions, security breaches, and unexpected behaviour.

Benefits : Enhance system performance, increase reliability, improve security, reduce maintenance costs, and prevent potential failures or downtime

OBJECTIVES



Research Questions

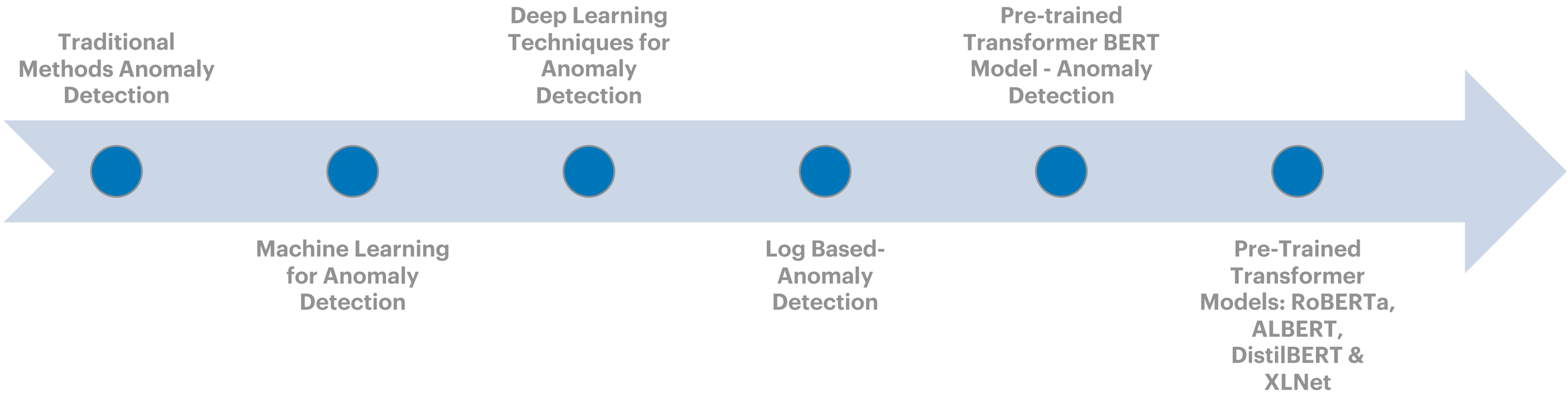
- How can a new approach using transformer method be identified to classify anomalies from system logs in real-time and address the issue of extracting meaningful information from unstructured logs?

- What are the drawbacks of recent log anomaly detection studies, such as concept drift, noise problems, and how can they be addressed to improve performance and reduce data instability and abnormal misjudgement?

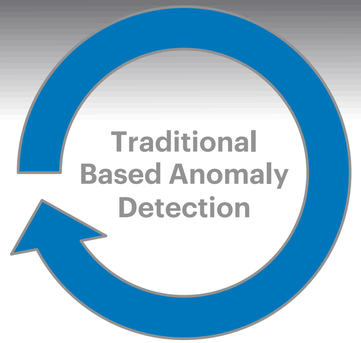
- What are the potential limitations of the proposed method, such as requiring a lot of computational resources and limited evaluation due to sensitive log and secure information, and how can they be overcome or mitigated to improve the effectiveness and generalizability of the proposed method?

Literature Review

The literature review encompasses a wide range of research endeavours in log anomaly detection, emphasizing the significance of automated methods for analysing system logs.



Literature Review - Continued



- Rules Based Anomaly Detection
- Statistical Based Anomaly Detection

Rule Based Anomaly Detection :

- Generate association rules from historical data to capture data relationships.
- Continuously monitor real-time data, labelling significant deviations from expected patterns as anomalies in rule-based detection.

Hela, S., Amel, B. and Badran, R., (2018) Early anomaly detection in smart home: A causal association rule-based approach. Artificial Intelligence in Medicine, 91, pp.57–71.

Statistical Based Anomaly Detection :

- Data-driven statistical techniques like Local Outlier Factor use attributes like mean and standard deviation to assess outlier-like behaviour in data.
- LOF assigns anomaly scores for unsupervised analysis, facilitating long-term monitoring of irregular data behaviour, aiding diverse anomaly detection types.

Jasra, S.K., Valentino, G., Muscat, A. and Camilleri, R., (2022) Hybrid Machine Learning–Statistical Method for Anomaly Detection in Flight Data. Applied Sciences (Switzerland), 1220.

Nooribakhsh, M. and Mollamotalebi, M., (2020) A review on statistical approaches for anomaly detection in DDoS attacks. Information Security Journal, .

Challenges :

Context-specific anomalies that differ from accepted statistical norms may be difficult for statistical tools to detect. Rules-based techniques, on the other-hand, have difficulty identifying unknown or unexpected anomalies that do not follow specified rules

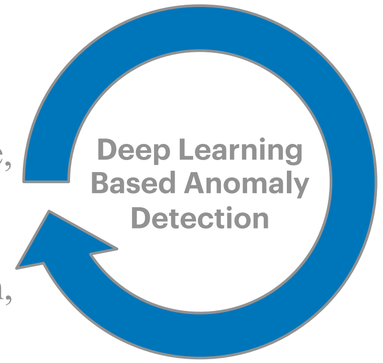
Literature Review - Continued

Machine Learning Based Anomaly Detection :

- Machine learning methods have gained attention for addressing anomaly detection challenges. SVMs and PCA are often used for efficiency and precision.
- Classification, ensemble, optimization, rule-based, clustering, and regression models, are employed alone or in hybrids for improved anomaly detection.

Nassif, A.B., Talib, M.A., Nasir, Q. and Dakalbab, F.M., (2021) Machine Learning for Anomaly Detection: A Systematic Review. IEEE Access, .

Challenges :



- Challenges in machine learning-based anomaly detection include high false positives, class dominance, noise sensitivity, and accuracy issues.
- Addressing these challenges requires ongoing research, alternative techniques, and the development of hybrid models in anomaly detection

Deep Learning Based Anomaly Detection :

- Tackling ML challenges: Deep learning, particularly LSTM models, were studied for anomaly detection. Researchers optimized LSTM with data pre-processing techniques.
- Network architectures: Fully Connected Networks, Variational Autoencoders, and Sequence-to-Sequence structures were explored with parameter tuning for better results.

.

Zhao, Z., Xu, C. and Li, B., (n.d.) A LSTM-Based Anomaly Detection Model for Log Analysis. [online] Available at: https://doi.org/10.1007/s11265-021-01644-4.

Malaiya, R.K., Kwon, D., Suh, S.C., Kim, H., Kim, I. and Kim, J., (2019) An Empirical Evaluation of Deep Learning for Network Anomaly Detection. IEEE Access, 7, pp.140806–140817.

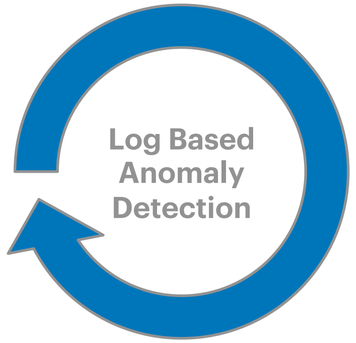
Challenges :

- LSTM and deep learning enhance anomaly detection accuracy but demand labelled data, computational resources, and interpretability improvements.
- Ongoing research strives to resolve these challenges and improve LSTM-based anomaly detection in log data

Literature Review - Continued

Log Based Anomaly Detection :

- We investigated various log parsers, including heuristic, frequency, clustering, and subsequence-based methods, on log-based anomaly detection.
- Surprisingly, high parsing accuracy alone didn't ensure better anomaly detection; heuristic-based parsers like Drain and IPLoM excelled by considering both accuracy and parsed event templates.

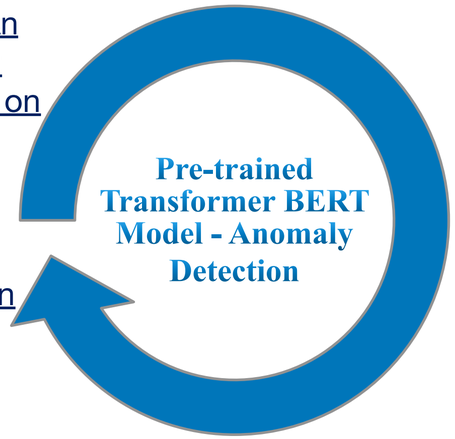


He, P., Zhu, J., Zheng, Z. and Lyu, M.R., (2017) Drain: An Online Log Parsing Approach with Fixed Depth Tree. In: Proceedings - 2017 IEEE 24th International Conference on Web Services, ICWS 2017. Institute of Electrical and Electronics Engineers Inc., pp.33–40.

Fu, Y., Yan, M., Xu, Z., Xia, X., Zhang, X. and Yang, D., (2023) An empirical study of the impact of log parsers on the performance of log-based anomaly detection. Empirical Software Engineering, 281.

Pretrained Transformer BERT Model-Anomaly Detection:

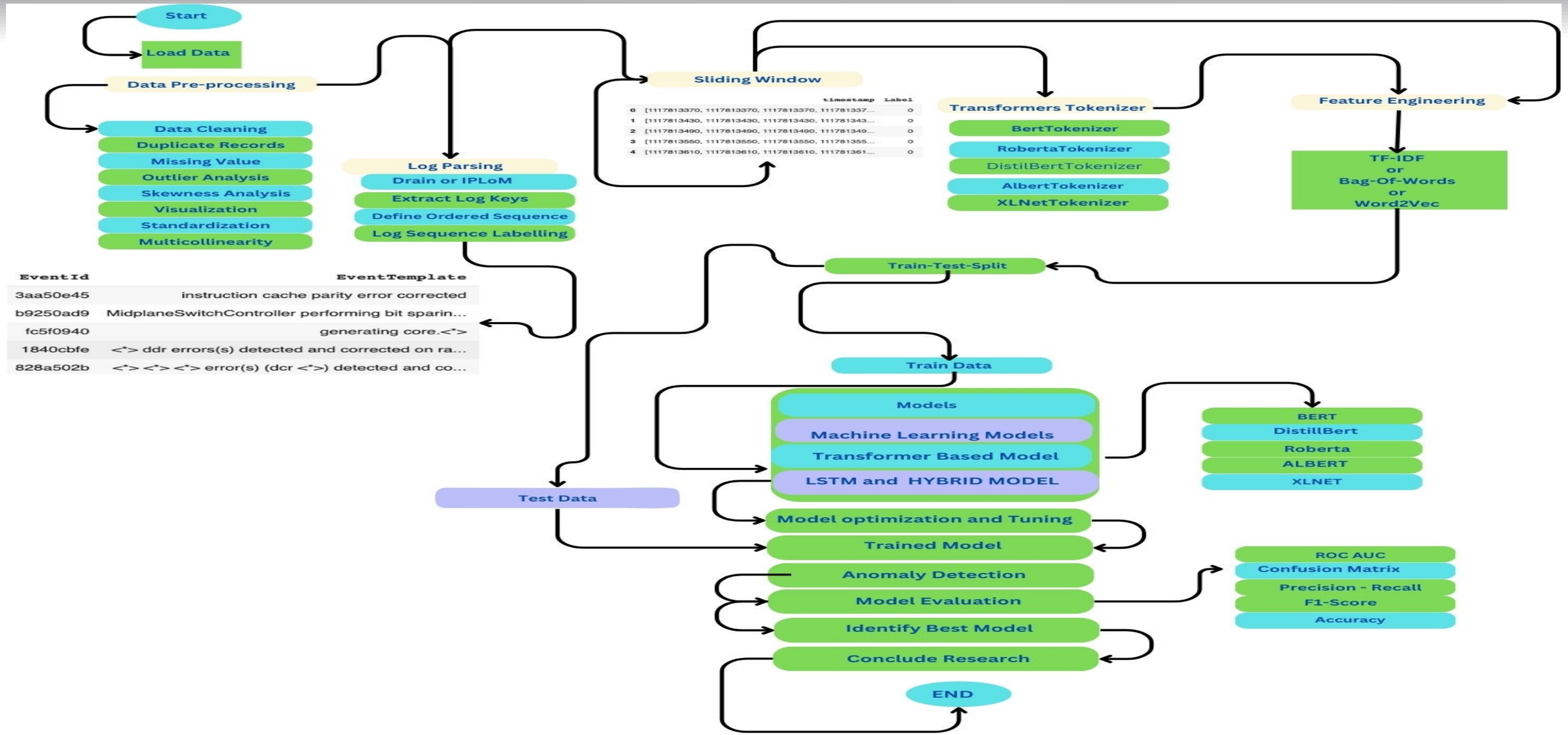
- Transformer-based models excel in capturing complex patterns in large datasets, making them ideal for sequential data.
- The BERT-Log method, using a pre-trained BERT model, proves highly effective in anomaly detection, particularly on HDFS and BGL datasets.



Wang, X., Cao, Q., Wang, Q., Cao, Z., Zhang, X. and Wang, P., (2022b) Robust log anomaly detection based on contrastive learning and multi-scale MASS. Journal of Supercomputing, 7816, pp.17491–17512.

Huang, S., Liu, Y., Fung, C., He, R., Zhao, Y., Yang, H. and Luan, Z., (2020a) HitAnomaly: Hierarchical Transformers for Anomaly Detection in System Log. IEEE Transactions on Network and Service Management, 174, pp.2064–2076.

Research Methodology



Exploratory Data Analysis

Missing Value Plot

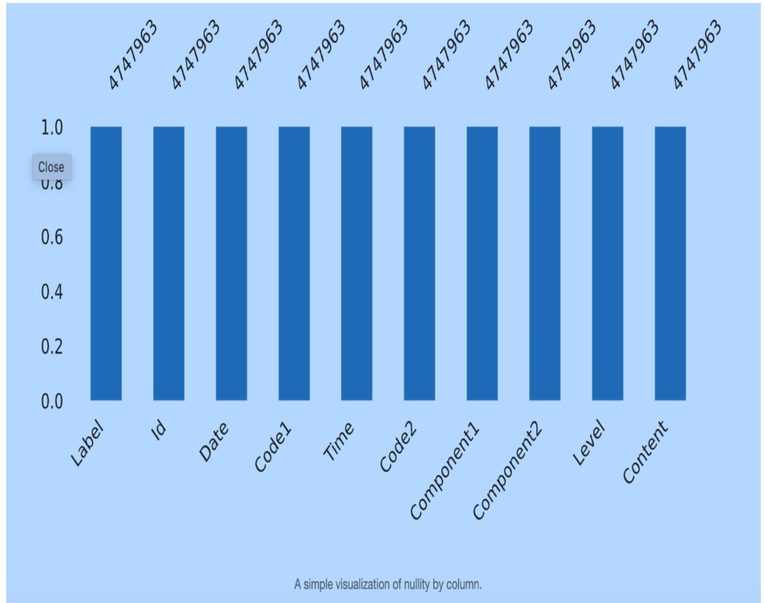


Figure 4.3.1 : Missing Values Visualization

Correlation Heatmap

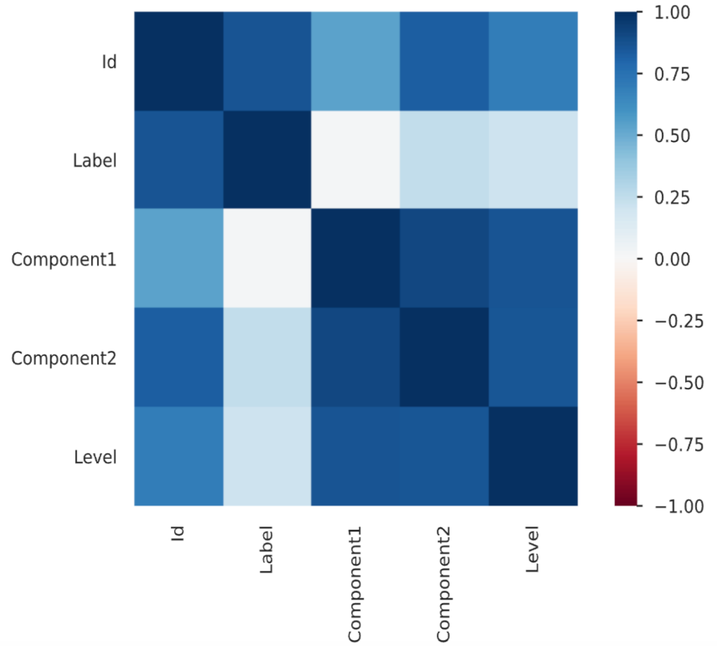


Figure 4.3.2 : Correlation Heatmap

Value	Count	Frequency (%)
-	4399503	92.7%
KERNDTLB	152734	3.2%
KERNSTOR	63491	1.3%
APPSEV	49651	1.0%
KERNMNTF	31531	0.7%
KERNTERM	23338	0.5%
KERNREC	6145	0.1%
APPREAD	5983	0.1%
KERNRTSP	3983	0.1%
APPRES	2370	< 0.1%
Other values (32)	9234	0.2%

Figure 4.3.3.1 : Label Categories Count and Frequency

Exploratory Data Analysis - Continued

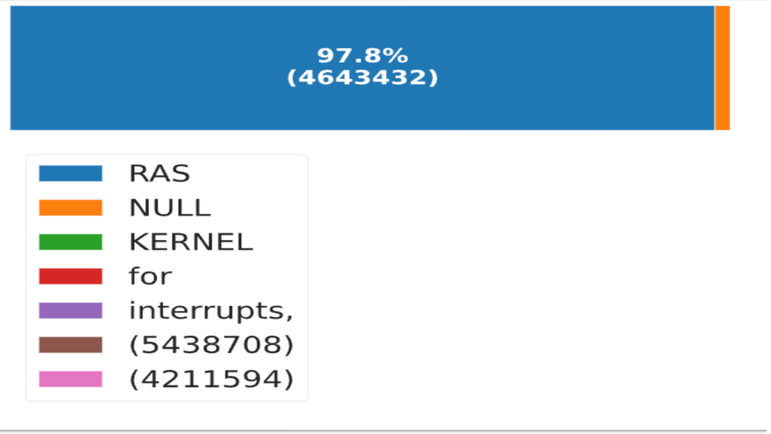


Figure 4.3.3.2 : Component Feature Common Value Plot

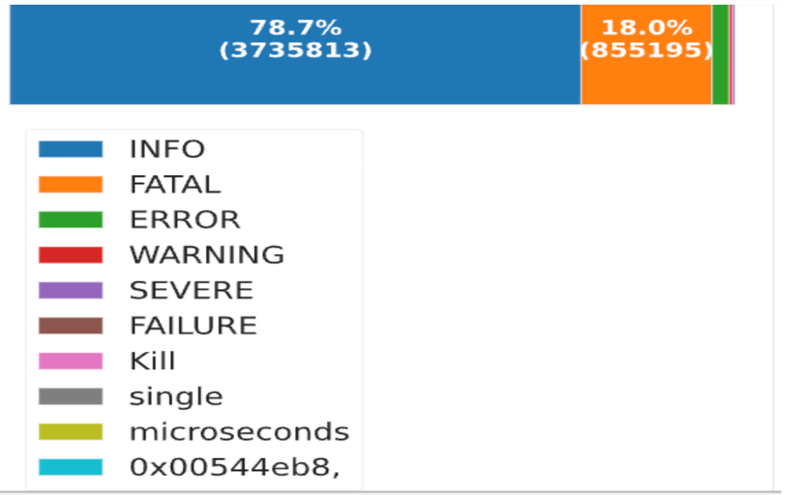


Figure 4.3.3.3 :Common value plot : Level

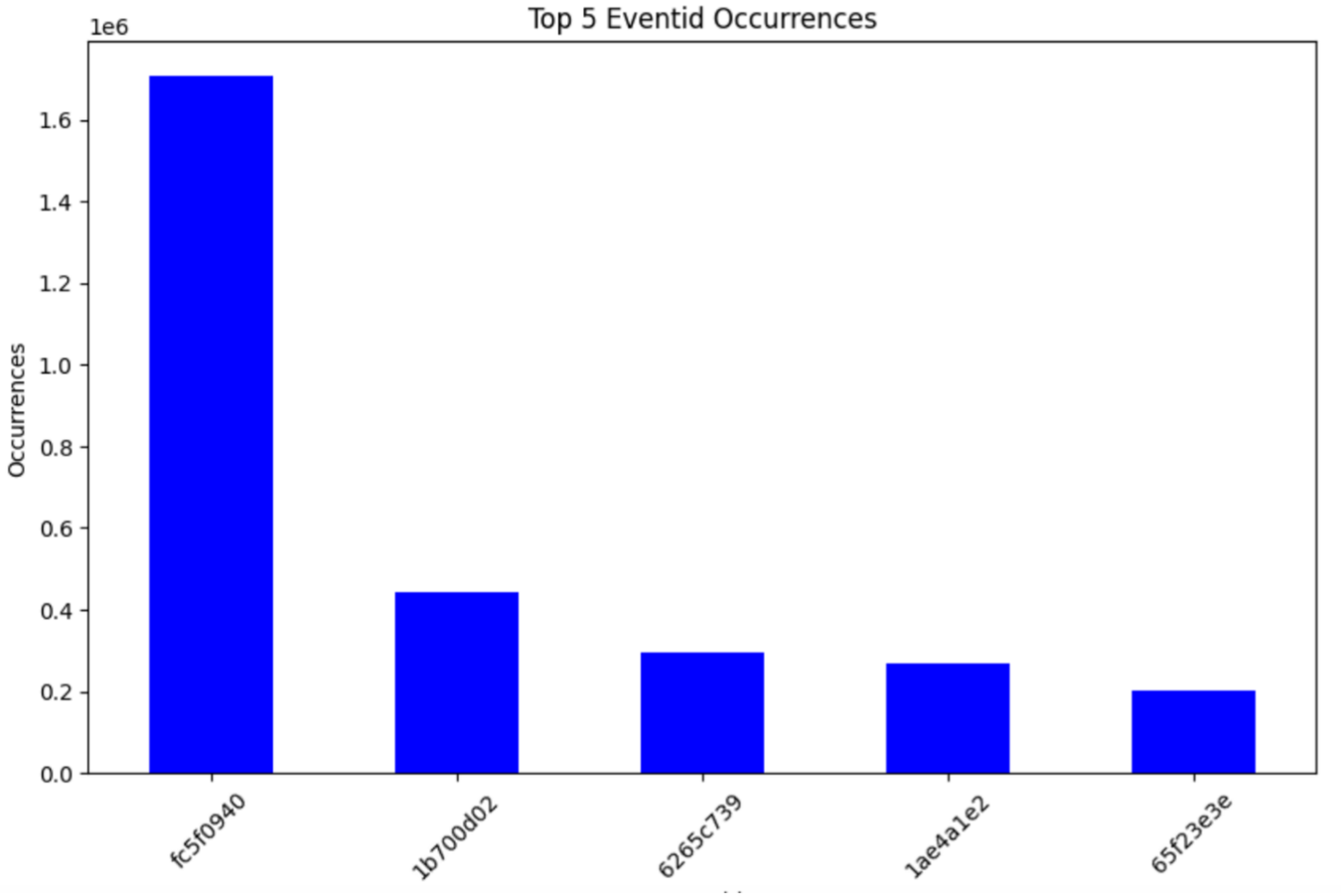


Figure 4.4.1.2 : Top 5 Event ID Occurrences

Feature Engineering

Log Parsing using Drain Algorithm

The approach involves utilizing the Drain algorithm for log parsing, which replaces specific patterns in log content, such as hexadecimal values, IP addresses, and numerical sequences, with a general symbol, denoted as "<*>". This process standardizes log entries, enhances readability, and filters out potentially sensitive or less relevant information, facilitating subsequent log analysis and troubleshooting

```
df_templates.head()
```

	EventId	EventTemplate	Occurrences
0	3aa50e45	instruction cache parity error corrected	105924
1	b9250ad9	MidplaneSwitchController performing bit sparin...	964
2	fc5f0940	generating core.<*>	1706751
3	1840cbfe	<*> ddr errors(s) detected and corrected on ra...	33939
4	828a502b	<*> <*> <*> error(s) (dcr <*>) detected and co...	5040

Figure 4.4.1.1 : Drain Log Parsing Template Results

Log sequencing using sliding window

The window size influences the amount of data considered for analysis, affecting the granularity of insights. The step size controls the overlap between consecutive windows, impacting how much information is shared between them. The timestamp and deltaT assist in aligning the sliding window with the temporal aspects of the data, ensuring accurate analysis and maintaining a contextually meaningful sequence of events.we have considered window_size = 300 and step size = 60 for log sequencing.

```
there are 37315 instances (sliding windows) in this dataset
```

	timestamp	Label	EventId	delta
0	[1117813370, 1117813370, 1117813370, 111781337...	0	[3aa50e45, 3aa50e45, 3aa50e45, 3aa50e45, 3aa50...	[0.0, 0.164068, 0.148025, 0.147847, 0.159012,
1	[1117813430, 1117813430, 1117813430, 111781343...	0	[3aa50e45, 3aa50e45, 3aa50e45, 3aa50e45, 3aa50...	[0.0, 0.154454, 0.15088, 0.137283, 0.165617, 6
2	[1117813490, 1117813490, 1117813490, 111781349...	0	[3aa50e45, 3aa50e45, 3aa50e45, 3aa50e45, 3aa50...	[0.0, 0.136003, 0.169535, 0.158248, 0.141224,
3	[1117813550, 1117813550, 1117813550, 111781355...	0	[3aa50e45, 3aa50e45, 3aa50e45, 3aa50e45, 3aa50...	[0.0, 0.158966, 0.136016, 0.160414, 0.148952,
4	[1117813610, 1117813610, 1117813610, 111781361...	0	[3aa50e45, 3aa50e45, 3aa50e45, 3aa50e45, 3aa50...	[0.0, 0.286636, 0.148804, 0.307791, 0.290853,

Figure 4.4.3.1 : Data Frame after sliding window execution

Model Experimentation

BASE LINE MODELS

- PCA model achieved an accuracy of 20.40%, its notably low precision and F1 score highlighted limitations in distinguishing genuine anomalies.
- The Isolation Forest displayed strong accuracy and precision, but its restricted recall affected its overall anomaly capture.
- The One- Class SVM model encountered challenges in both accurate classifications and comprehensive anomaly detection.
- The LogClustering model, with high precision and respectable recall, showcased adeptness in capturing real anomalies.
- Notably, the LogClustering models stood out with commendable performance, emphasizing the importance of precision-recall balance also reliant on specific data characteristics and parameter choices, warranting further exploration for enhanced performance across diverse data scenarios.

PCA

Metric	Value
Confusion Matrix	
TP (True Positives)	1194
FP (False Positives)	11868
TN (True Negatives)	1851
FN (False Negatives)	14
Precision	9.141%
Recall	98.841%
F1-Measure	16.734%
Accuracy Score	0.20399276478

Isolation Forest

Metric	Value
Confusion Matrix	
True Positives (TP)	170
False Positives (FP)	0
True Negatives (TN)	13,719
False Negatives (FN)	1,038
Precision	100.000%
Recall	14.073%
F1-measure	24.673%
Accuracy Score	93.05%

One-Class SVM

	Value
Confusion Matrix	
True Positives (TP)	144
False Positives (FP)	13,719
True Negatives (TN)	0
False Negatives (FN)	1,064
Precision	1.039
Recall	11.921
F1-Measure	1.911
Accuracy Score	0.0096

LogClustering

	Value
Confusion Matrix	
True Positives	798
False Positives	29
True Negatives	13,690
False Negatives	410
Precision	96.493%
Recall	66.060%
F1-measure	78.427%
Accuracy Score	97.059%

Model Experimentation - Continued

TRANSFORMER MODELS

Bert_base_uncased, Tuned Bert, DistilBERT, RoBERTa, Albert, and XLNet, has impressively demonstrated their prowess in forecasting anomalies

The DistilBERT model emerges as an exceptional standout, showcasing robust performance across a spectrum of assessment metrics.

- Boasting an average accuracy of 98.00%, the model adeptly labels roughly 98% of test instances with precision.
- Its precision score of 99.00% underscores its adeptness in curbing false positives, while an average recall of 81.00% underscores its ability to identify 81% of genuine positive cases.
- By harmonizing precision and recall, the model achieves an average F1-score of 88.00%, signifying its holistic efficacy.
- It is well-suited for applications entailing real-time or resource-constrained processing.

Input Parameters:

- batch_size = 16 , Num_labels = 2: Criterion = nn.CrossEntropyLoss(). ,Optimizer = torch.optim.Adam(bert_base_model.parameters(), lr=1e-5): Max_seq_length = 512:

Model Structure:

- An initial fully connected layer featuring 256 units, followed by a ReLU activation.
- A dropout layer with a 20% dropout rate.
- A final fully connected layer, mapped to the specified number of labels.

BERT - BASE

Metric	Score
Test Average Accuracy	0.9800
Test Average Precision	0.9900
Test Average Recall	0.8000
Test Average F1-Score	0.8800

Tuned BERT

Metric	Value
Training Loss	0.0756
Test Accuracy	98.00%
Test Precision	97.00%
Test Recall	82.00%
Test F1-Score	88.00%

DistilBERT

Metric	Value
Training Loss	0.0752
Test Accuracy	98.00 %
Test Precision	99.00 %
Test Recall	81.00 %
Test F1-Score	88.00 %

Roberta

Metric	Value
Training Loss	0.0710
Test Accuracy	98.00 %
Test Precision	94.00 %
Test Recall	77.00 %
Test F1-Score	84.00 %

ALBERT

Metric	Value
Training Loss	0.0609
Test Accuracy	98.00 %
Test Precision	98.00 %
Test Recall	81.00 %
Test F1-Score	88.00 %

XLNET

Metric	Value
Training Loss	0.0668
Test Accuracy	98.00 %
Test Precision	99.00 %
Test Recall	79.00 %
Test F1-Score	87.00 %

Model Experimentation - Continued

LSTM MODELS

LSTM , Tuned-LSTM , BI-LSTM demonstrates strong performance across a variety of assessment metrics ,

The BiLSTM model excels across diverse evaluation metrics.

Interpretation:

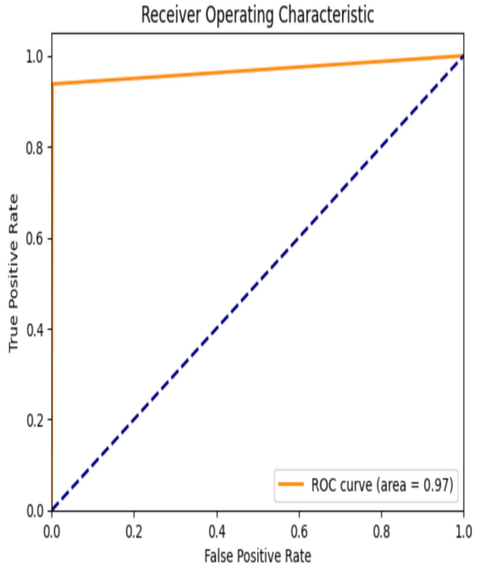
•Impressive test accuracy of 99.33% signifies precise classification.

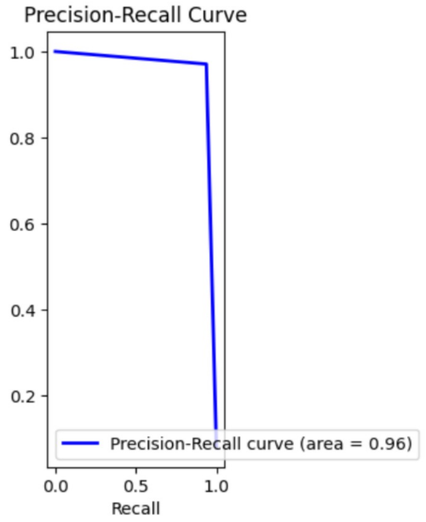
•With an average precision of 97.68%, the model aptly manages false positive instances.

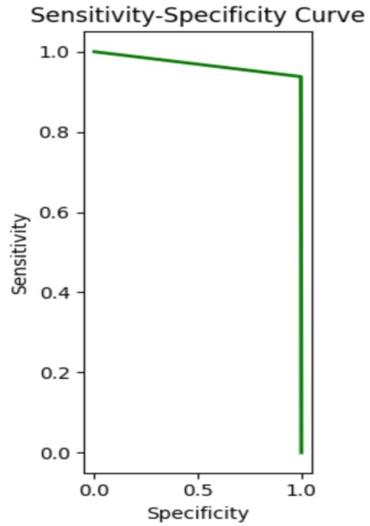
•An average recall of 93.96% highlights the model's efficiency in capturing genuine positive instances.

•The average F1-score of 95.78% adeptly balances precision and recall, indicating Comprehensive performance.

Bi –LSTM MODEL







Input Parameters

- **input_size** = 1241
- **hidden_size** = 256
- **num_layers** = 2
- **num_classes** = 2
- **dropout** = 0.2
- **batch_size** = 32
- **learning_rate** = 0.001
- **num_epochs** = 20

Metric	Value
Training Loss	0.0347
Test Accuracy	99.33 %
Test Precision	97.68 %
Test Recall	93.96 %
Test F1-Score	95.78 %

Model Experimentation - Continued

Hybrid MODELS – LSTM + BERT

Tailoring the model to specific task nuances and data attributes is recommended for optimal outcomes.

Model Architecture

The Hybrid Model seamlessly combines the strengths of BERT and LSTM to enhance anomaly detection. The architecture integrates BERT for contextual embeddings, followed by an LSTM layer to capture temporal patterns, and a fully connected layer for classification.

Input Parameters

•The chosen BERT model is 'bert-base-uncased' for leveraging pre-trained capabilities.
•The LSTM input size is tailored to fit the data at hand (1241 in this case).
•A hidden size of 256 is selected to determine the LSTM's hidden state dimensions.
•Two LSTM layers are stacked to optimize the temporal learning process.
•The classification task involves two classes.
•Data is processed in batches of 100 for efficiency.
•Learning occurs with a rate of 0.001.
•The model undergoes training over 5 epochs.

Evaluation Matrix

Metric	Value
Training Loss	0.4045
Test Accuracy	93.92 %
Test Precision	88.57 %
Test Recall	25.40 %
Test F1-Score	38.12 %

Model Evaluation and Summary

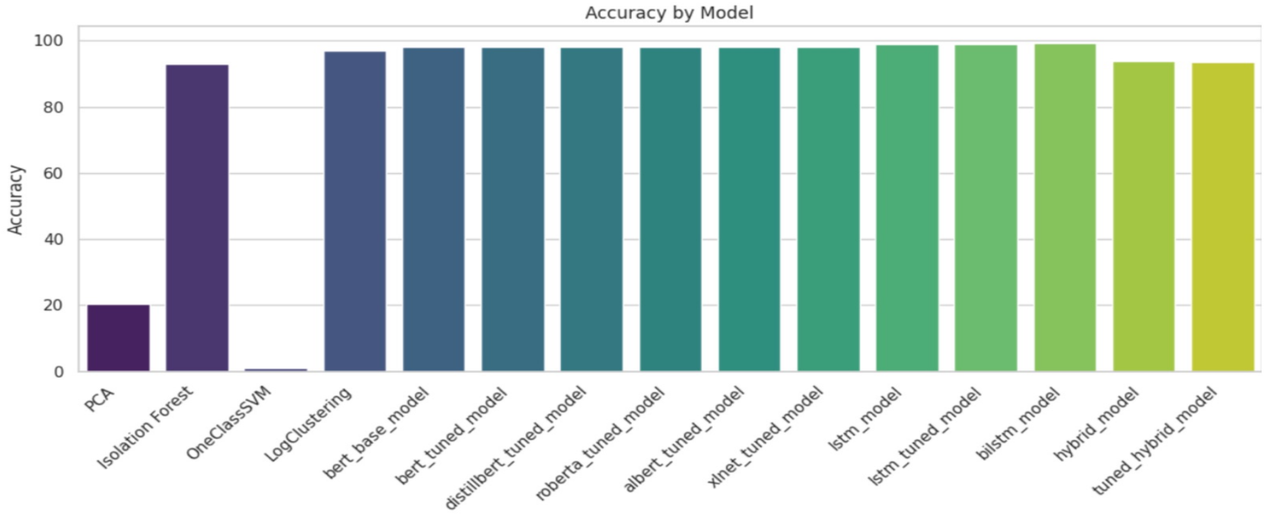
Metrics Comparison

```
df_metrics # Below Table values are in Percentage
```

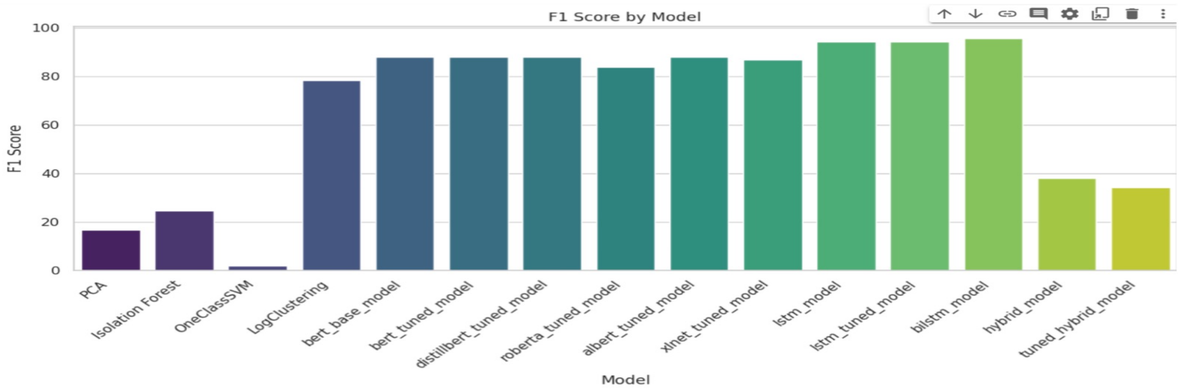
	Model	Accuracy	Precision	Recall	F1 Score
0	PCA	20.3993	9.141	98.8411	16.7344
1	Isolation Forest	93.0462	100.0	14.0728	24.6734
2	OneClassSVM	0.9647	1.0387	11.9205	1.911
3	LogClustering	97.0590	96.4933	66.0596	78.4275
4	bert_base_model	98.0000	99.0000	80.0000	88.0000
5	bert_tuned_model	98.0000	97.0000	82.0000	88.0000
6	distillbert_tuned_model	98.0000	99.0000	81.0000	88.0000
7	roberta_tuned_model	98.0000	94.0000	77.0000	84.0000
8	albert_tuned_model	98.0000	98.0000	81.0000	88.0000
9	xlnet_tuned_model	98.0000	99.0000	79.0000	87.0000
10	lstm_model	99.0688	94.0643	94.4536	94.2586
11	lstm_tuned_model	99.0889	95.0420	93.6258	94.3286
12	bilstm_model	99.3301	97.6764	93.9570	95.7806
13	hybrid_model	93.9153	88.5667	25.3977	38.1159
14	tuned_hybrid_model	93.6553	85.1444	22.5141	34.2869

Figure 5.3.1 : Model Metrics Evaluation Comparison

- Comparative analysis of all experimented models accuracy score



- Comparative analysis of all experimented models F1 score



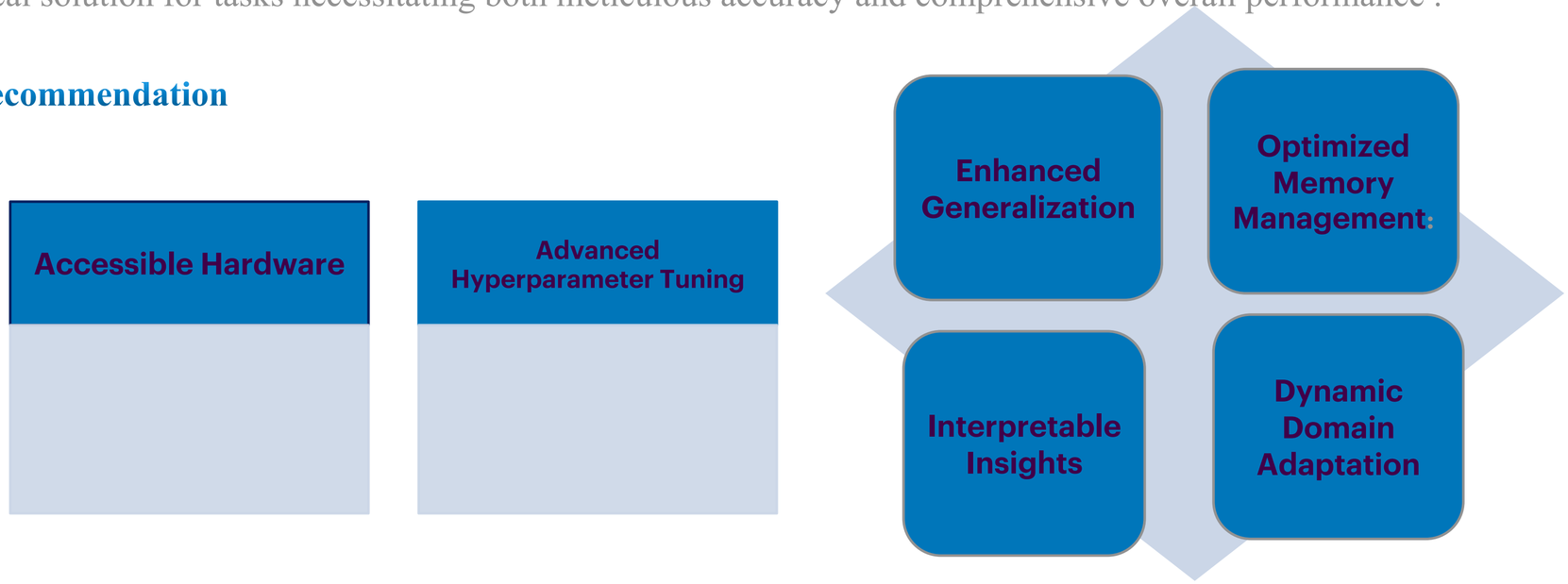
Conclusion and Future Recommendation

Conclusion

- The research extensively explored various anomaly detection models, including baseline ones like PCA, Isolation Forest, OneClassSVM, and LogClustering, as well as advanced Transformer-based models such as Bert_base_uncased, Tuned Bert, DistilBERT, RoBERTa, Albert, and XLNet, along with LSTM models and a Hybrid Model.

- In summary, each recommended model brings distinct strengths to the table, catering to diverse anomaly detection requirements. The choice among these models hinges on task-specific characteristics, the desired precision-recall equilibrium, and available resources. Considering the metrics and the model's performance balance, the BiLSTM model emerges as the quintessential choice for predictive tasks. With its exceptional accuracy, precision, recall, and F1-score, it embodies the multifaceted capacities crucial for real-world applications, positioning it as an ideal solution for tasks necessitating both meticulous accuracy and comprehensive overall performance .

Future Recommendation



Limitations