International Journal of Science and Engineering Applications
Volume 14-Issue 01, 65 - 72, 2025, ISSN:- 2319 - 7560
DOI:10.7753/IJSEA1401.1015

A Novel Deep Learning Model for Fault Detection in Power Transformers

Nagaraju Brahmanapally
Student, Department of CS
University of West Florida
Pensacola, FL 32514, USA

Drake Fulton
Student, Department of ECE
University of West Florida
Pensacola, FL 32514, USA

Dr. Bhuvana Ramachandran
Professor, Department of ECE
University of West Florida
Pensacola, FL 32514, USA

Abstract: Power transformers are vital in ensuring the reliability of electrical power systems, necessitating accurate fault classification for their efficient operation. This research evaluates a novel Transformer Deep Learning model architecture for fault classification using dissolved gas analysis (DGA) data, leveraging feature engineering and an over-sampling technique to address high-dimensionality and class imbalance challenges. The model demonstrated substantial accuracy improvements across datasets of varying sizes and preprocessing stages, particularly with SMOTE-enhanced data. These findings underscore the effectiveness of Transformer deep learning architectures in advancing the state-of-the-art in fault classification for power transformer systems.

Keywords: Power Transformer Fault Classification, Dissolved Gas Analysis , Transformer Deep Learning Model, Fault Type Prediction, Gas Concentration Analysis, Feature Engineering, Synthetic Minority Over-sampling Technique, Imbalanced Dataset Handling

1. INTRODUCTION

Power transformers are critical components in electrical power systems, responsible for the efficient transmission and distribution of electricity. The reliability of these transformers is paramount to ensuring a stable and uninterrupted power supply. However, transformers are prone to various types of faults, often indicated by the presence of dissolved gases in the transformer oil. The accurate classification of these faults based on gas concentration levels is crucial for timely maintenance and prevention of catastrophic failures. Traditional methods for fault diagnosis, such as the Duval Triangle method and Key Gas Analysis, have limitations in handling the complex, high-dimensional data typically encountered in modern power systems. This has led to the exploration of advanced machine learning and deep learning techniques for more accurate and automated fault classification [1].

In recent years, Transformer models have gained significant attention in various domains, including natural language processing and time-series analysis, due to their ability to capture long-range dependencies and complex relationships within data [2]. The self-attention mechanism, which is central to Transformer architectures, enables these models to weigh the importance of different input features dynamically, making them highly effective in tasks requiring nuanced understanding of input data. In the context of power transformer fault diagnosis, the application of Transformer models is relatively novel. Their potential to handle high-dimensional data and learn intricate patterns makes them promising candidates for improving fault classification accuracy, especially when dealing with data that exhibits significant variability and noise.

Recent research in the field of power transformer fault diagnosis has begun to explore the use of advanced deep learning techniques. For example, Zhi Li et al. [3] proposed a fault diagnosis technique based on Long Short-Term Memory (LSTM) neural networks combined with dissolved gas analysis (DGA). Their study, which analyzed 240 samples, demonstrated that the LSTM model achieved superior diagnostic accuracy compared to traditional neural networks. This underscores the potential of deep learning models to improve fault diagnosis in power transformers. Despite these advancements, there remains a gap in the application of Transformer models specifically for predicting power transformer fault types, suggesting a novel direction for future research [4].

In addition to the challenges posed by high-dimensional data, another critical issue in transformer fault diagnosis is the class imbalance often present in the data. Certain fault types may occur less frequently, leading to a skewed distribution that can bias machine learning models towards the more common classes. To address this issue, Synthetic Minority Over-sampling Technique (SMOTE) has been widely adopted as a data augmentation strategy. SMOTE generates synthetic samples for the minority class by interpolating between existing samples, thereby balancing the class distribution and enabling the model to learn from a more representative dataset [5]. The effectiveness of SMOTE has been demonstrated in various domains, including medical diagnosis, fraud detection, and power systems, where it has been used to enhance the performance of classifiers in imbalanced datasets [6,7]. In the power transformer domain, SMOTE, combined with feature engineering, can significantly improve the model's ability to correctly identify rare but critical fault types.

Several studies have applied Transformer models and SMOTE in different domains with positive outcomes. For instance, Transformer models have been used in the healthcare sector to predict patient outcomes based on time-series data, demonstrating superior accuracy compared to traditional recurrent neural networks (RNNs) [8]. Similarly, SMOTE has been successfully employed in fraud detection tasks to address the issue of imbalanced datasets, leading to more accurate and reliable predictions [6]. Despite these successes, there has been limited exploration of these techniques in the power transformer fault diagnosis domain. This study aims to bridge this gap by evaluating the performance of a Transformer-based model on a dataset of gas concentrations, with a particular focus on the impact of SMOTE and feature engineering on classification accuracy.

By leveraging the strengths of Transformer models and addressing the challenges of imbalanced datasets through SMOTE, this research seeks to advance the state-of-the-art in power transformer fault classification. The results presented in this paper not only demonstrate the efficacy of these methods in this domain but also provide insights into their potential application in other critical infrastructure systems where fault diagnosis is essential for maintaining operational reliability.

2. DATASETS

To implement the proposed deep learning model for identifying and classifying faults in transformers, three datasets were collected and processed. The small dataset was obtained from [9], manually entered in a spreadsheet, and saved as a CSV file. This dataset had no missing values and was 100% complete. The medium dataset, collected from [10], contained several missing values for gas concentrations. To ensure data consistency and prevent skewed results, rows with missing values were removed before the dataset was used in the model. This dataset was also saved as a CSV file. Additionally, the large dataset was sourced from [11] and combined with data from [10] to create the most extensive dataset. To address any skewed distributions and optimize the performance of the transformer model, all datasets underwent standardization. This process, which transforms data to have a zero mean and unit standard deviation, enhances the effectiveness of the algorithms.

2.1 Dataset Overview and Preprocessing

2.1.1 Small Dataset

The small dataset comprises 70 samples, each containing six key gas concentration features: hydrogen (H2), methane (CH4), acetylene (C2H2), ethylene (C2H4), ethane (C2H6), and carbon monoxide (CO), all measured in parts per million (ppm). Additionally, the dataset includes a target variable labeled "Fault," which is categorized into four distinct classes: "Thermal," "High Discharge," "Low Discharge," and "No Fault". These fault classes represent the specific fault types to be predicted. The distribution of these fault classes within the dataset is presented in Figure 1.

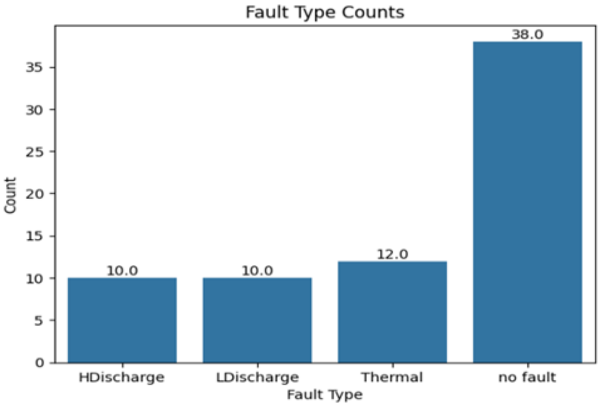

Figure 1: Fault distribution in small dataset

2.1.2 Medium Dataset

The medium dataset initially comprised 151 data points, each containing seven gas concentration features: hydrogen (H2), methane (CH4), acetylene (C2H2), ethylene (C2H4), ethane (C2H6), carbon monoxide (CO), and carbon dioxide (CO2), measured in parts per million (ppm), along with a target variable labeled "Fault". The fault types in this dataset included: 'D1' (Low Energy Discharge), 'D2' (High Energy Discharge), 'None' (No fault), 'HThermal' (High Thermal—thermal faults exceeding 700oC, as determined by equipment inspection), 'LThermal' (Low Thermal—thermal faults below 700oC, as determined by equipment inspection), and 'PD' (Partial Discharge). However, 37 data points had missing values for these gas features. To prevent these missing values from negatively impacting the algorithm's performance and introducing bias, these data points were excluded from the dataset. This likely resulted from unrecorded measurements. After cleaning, the final dataset consisted of 114 complete data points, which were used for subsequent analysis. The distribution of fault types is depicted in Figure 2.

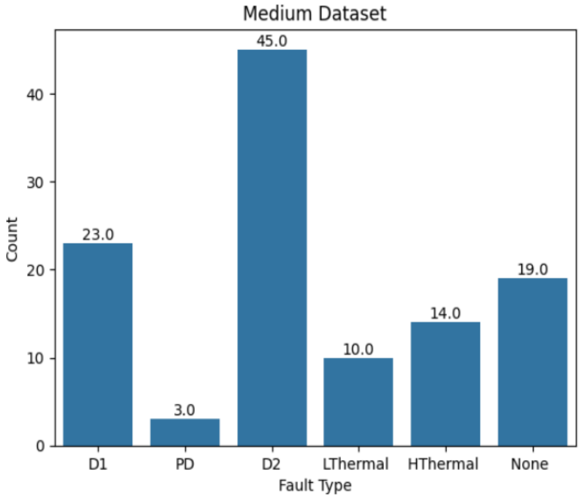

Figure 2: Fault distribution in medium dataset

2.1.3 Large Dataset

The large dataset originally consisted of 231 data points, each containing five gas concentration features: hydrogen (H2), methane (CH4), acetylene (C2H2), ethylene (C2H4), ethane (C2H6), all measured in parts per million (ppm), along with a target variable labeled "Fault." However, 18 data points had missing values for these gas features. To ensure the accuracy of the algorithm and prevent data bias, these incomplete data points were excluded, likely due to unrecorded measurements. After this data cleaning process, the final dataset comprised 213 complete data points, suitable for further analysis. The dataset includes nine distinct fault types, as shown in Table 1. The distribution of these fault types is depicted in Figure 3.

Table 1: Faults and number association

Fault Type	Number
Partial Discharge	0
Spark Discharge	1
Arc Discharge	2
High -temperature Overheating	3
Middle -Temperature	4
Low -Temperature Overheating	5
Low/Middle -Temperature	6
High Energy Discharge	7
No Fault	8

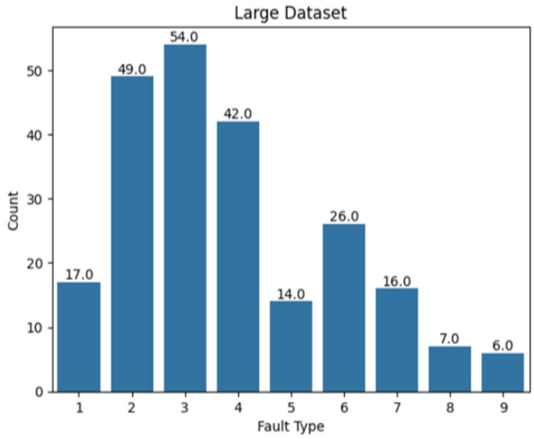

Figure 3: Fault distribution in large dataset

2.2 Feature Engineering: Unveiling Critical Insights

To enhance the performance of the transformer deep learning model, feature engineering was conducted based on a thorough correlation analysis of the gas concentration features within each dataset. Since the gas measurements varied across the datasets, the features generated through the feature engineering process were unique to each case. The correlation matrix heat maps revealed varying degrees of correlation between the features. Features with strong positive correlations and weak negative correlations were used to derive new features through ratio calculations. This strategy was designed to leverage the inherent relationships between gas concentrations while minimizing the influence of features with weaker or opposing trends. Specifically, features with moderate to high positive correlations (greater than 0.7) and weak negative correlations (around -0.1) were selected to create these new ratio-based features. Consistent patterns of correlations were observed across all three datasets.

2.2.1 Small Dataset

In the small dataset, only ratios derived from highly positive correlations were identified and utilized to generate additional features. Six key correlations were selected to create new ratio-based features:

• H2:C2H6 Ratio: The strong positive correlation (corr = 0.92) between Hydrogen(H2) and Ethane (C2H6) suggests a close relationship in their concentrations under fault conditions. This ratio captures and leverages this inherent link.

• H2:CO Ratio: A strong positive correlation (corr = 0.78) exists between Hydrogen (H2) and Carbon (CO) indicating that a higher H2:CO ratio may reflect similar trends in fault-related gas emissions.

• CH4:C2H6 Ratio: The strong positive correlation (corr = 0.87) between Methane (CH4) and Ethane (C2H6) reflect their tendency to vary together, making this ratio a valuable feature for distinguishing fault conditions.

• CH4:CO Ratio: The positive correlation (corr = 0.79) between Methane (CH4) and Carbon (CO) suggest that their concentrations rise and fall together, potentially offering predictive insights through the CH4:CO ration.

• C2H2:C2H4 Ratio: The strong positive correlation (corr = 0.79) between Acetylene (C2H2) and Ethylene (C2H4) highlights their mutual response to fault conditions, making this ratio a meaningful feature for fault classification.

• C2H6:CO Ratio: The strong positive correlation (corr = 0.80) between Ethane (C2H6) and Carbon (CO) further emphasizes the relationship between these gases, allowing this ratio to capture relevant fault-related interactions.

The correlation heatmap for the small dataset is represented in Figure 4.

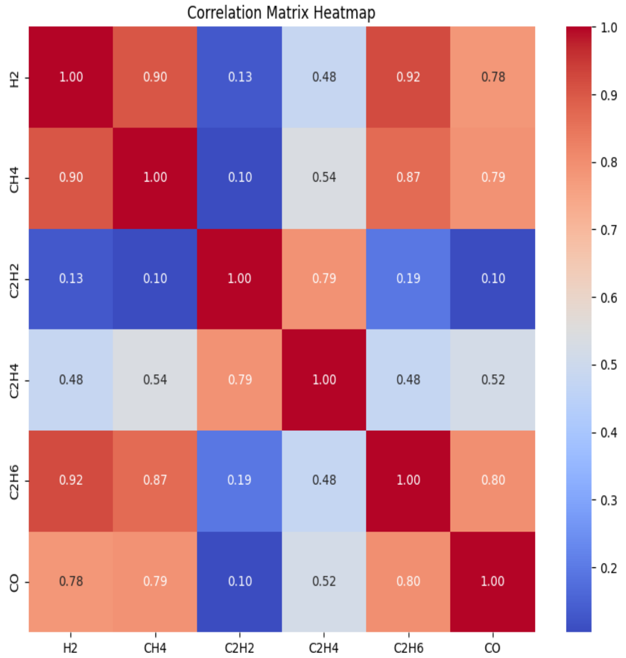

Figure 4: Correlation heat map of the small dataset

2.2.2 Medium Dataset

In the medium dataset, both ratios with high positive correlations and those with weak negative correlations were identified and used to generate additional features. Five key correlations were selected for feature creation:

• H2:CO2 Ratio: This ratio captures the relative concentration of Hydrogen (H2) to Carbon Dioxide (CO2). Although the correlation is weakly negative (corr = -0.06), a higher H2:CO2 ratio may still be indicative of specific fault types.

• CH4:C2H4 Ratio: The strong positive correlation (corr = 0.85) between Methane (CH4) and Ethylene (C2H4) highlights their potential interdependence during fault conditions, making this ratio an informative feature for fault prediction.

• C2H2:CO2 Ratio: Similar to the H2:CO2 ratio, this feature (corr = -0.09) represents the relative concentration of Acetylene (C2H2) to Carbon Dioxide (CO2). Despite the weak negative correlation, this ratio could provide subtle insights into fault characteristics.

• C2H4:C2H6 Ratio: The high positive correlation (corr = 0.76) between Ethylene (C2H4) and Ethane (C2H6) indicates that their concentrations tend to increase or decrease together, providing valuable information for the model through the C2H4: C2H6 ratio.

• CO: CO2 Ratio: The positive correlation (corr = 0.70) between Carbon Monoxide (CO) and CO2 suggests a co-dependent behavior of these gases under transformer fault conditions, making the CO:CO2 ratio a key feature in the dataset.

The correlation heatmap for the medium dataset is illustrated in Figure 5.

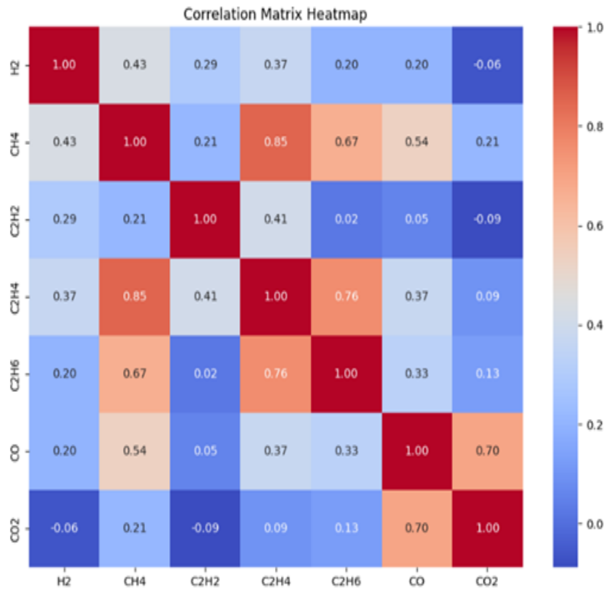

Figure 5: Correlation heat map of the medium dataset

2.2.3 *Large Dataset*

In the large dataset, only ratios reflecting high positive correlations were identified. Three key correlations were established to create additional features:

- CH4:C2H6 Ratio: This ratio represents the relative concentration of Methane (CH4) to Ethane (C2H6). The strong positive correlation (corr = 0.78) suggests that an increased CH4:C2H6 ratio may indicate similar trends in gas emissions.
- CH4:C2H4 Ratio: A strong positive correlation (corr = 0.87) exists between Methane (CH4) and Ethylene (C2H4), indicating a potential connection in their concentrations during fault conditions. This ratio serves to leverage the inherent relationship within the dataset.
- C2H4:C2H6 Ratio: The very high positive correlation (corr = 0.92) between Ethylene (C2H4) and Ethane (C2H6) signifies that their concentrations tend to rise and fall in tandem under specific fault conditions, providing valuable insights for the model.

The correlation heatmap for the large dataset is presented in Figure 6.

2.3 Data Augmentation and Balancing with SMOTE

2.3.1 *Synthetic Minority Oversampling Technique (SMOTE)*

The Synthetic Minority Over-Sampling Technique (SMOTE) is a widely used approach for addressing class imbalance in machine learning tasks. Proposed by Chawla et al. [12] in their paper "SMOTE: Synthetic Minority Over-sampling Technique," SMOTE generates synthetic examples of the minority class to balance the distribution of classes in the training dataset. This method aims to enhance the model's ability to learn from underrepresented classes by providing a more even representation of all classes [12].

SMOTE operates by creating synthetic samples in the feature space rather than simply duplicating existing samples. It selects samples that are close in the feature space and generates new samples along the line segments connecting them. This process effectively increases the density of the minority class and improves the model's performance in detecting and classifying underrepresented fault types [12].

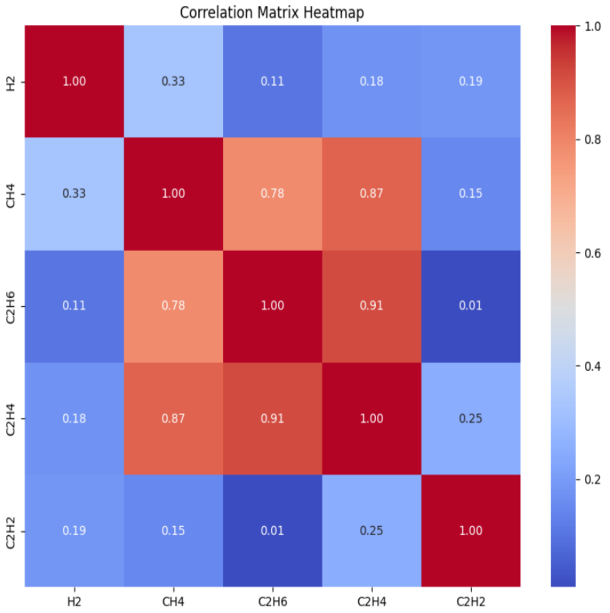

Figure 6: Correlation heat map of the large dataset

The application of SMOTE spans various domains, including medical diagnosis, fraud detection, and industrial equipment fault prediction. In medical diagnostics, SMOTE has been used to improve the detection of rare diseases and anomalies by generating synthetic patient data [13]. In fraud detection, SMOTE helps in identifying fraudulent transactions in imbalanced financial datasets [14]. For industrial applications, such as power transformer fault prediction, SMOTE addresses the challenge of class imbalance by enhancing the model's ability to detect and predict rare but critical fault types [15]. By incorporating SMOTE, predictive models can achieve better performance and more reliable predictions in scenarios where class imbalance is a significant issue.

2.3.2 *Data Augmentation with SMOTE*

Across all datasets, instances of class imbalance were observed. Such imbalances can lead to inconsistencies during the training of the proposed algorithm, as unbalanced data may result in inaccurate predictions due to the dominance of oversampled classes. Each of the three datasets contained minority classes that could contribute to misclassifications. To address these imbalances, the Synthetic Minority Oversampling Technique (SMOTE) was employed. This technique was applied to all three datasets. Figures 7, 8, and 9 illustrate the distributions of fault types following the application of SMOTE, which equalized the number of samples for each fault type, resulting in a balanced representation across all classes.

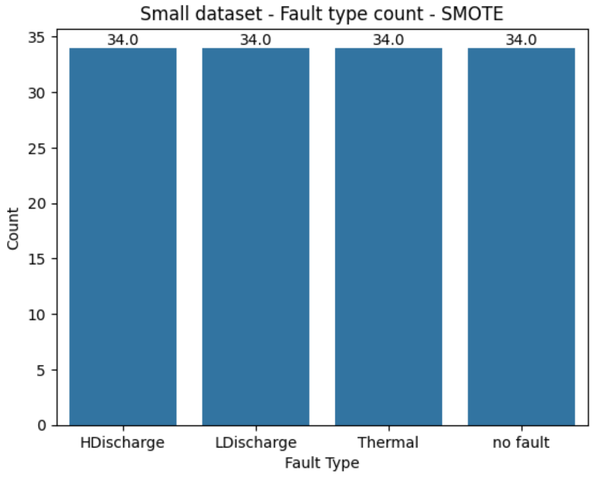

Figure 7: Fault distribution in small dataset after Feature engineering & SMOTE

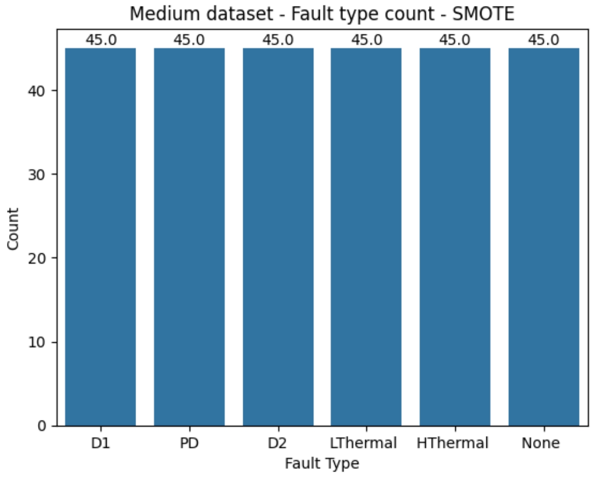

Figure 8: Fault distribution in medium dataset after SMOTE

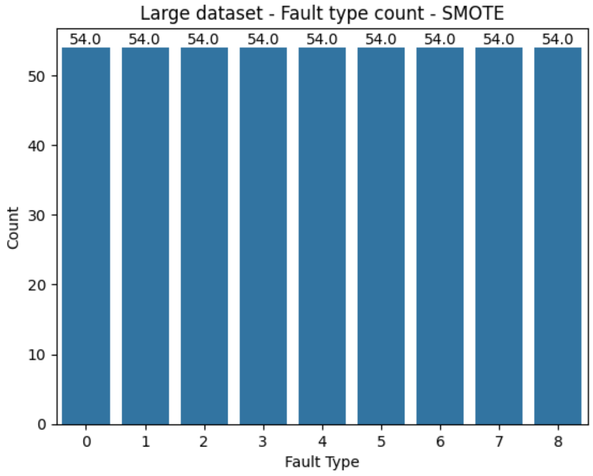

Figure 9: Fault distribution in large dataset after SMOTE

3. DEEP LEARNING TRANSFORMER MODEL

Transformer models represent a significant advancement in the field of deep learning, particularly in handling sequence-to-sequence tasks. Introduced by Vaswani et al. [2] in their seminal paper "Attention Is All You Need," transformers utilize a self-attention mechanism that enables the model to weigh the importance of different elements in the input sequence, irrespective of their position. This mechanism is crucial for capturing long-range dependencies and contextual relationships within data [2].

The architecture of transformers consists of encoder and decoder layers, each equipped with multi-head self-attention and feed-forward neural networks. This design allows transformers to process sequences in parallel, significantly improving efficiency compared to previous sequential models like RNNs and LSTMs [2]. Transformers have achieved state-of-the-art results in various domains, including natural language processing (NLP) and computer vision. In NLP, models like BERT (Bidirectional Encoder Representations from Transformers) and GPT (Generative Pre-trained Transformer) have set new benchmarks in tasks such as text classification, translation, and summarization [16][17].

In the context of time-series and predictive maintenance, transformers offer promising capabilities. Their ability to handle complex temporal dependencies makes them suitable for analyzing sensor data and predicting faults in industrial systems. Recent research has demonstrated the effectiveness of transformers in fault detection and prediction for various types of equipment, including power transformers [18]. The versatility of transformer model (shown in Figure 10) in capturing intricate patterns and relationships in data positions them as a powerful tool for improving fault prediction accuracy.

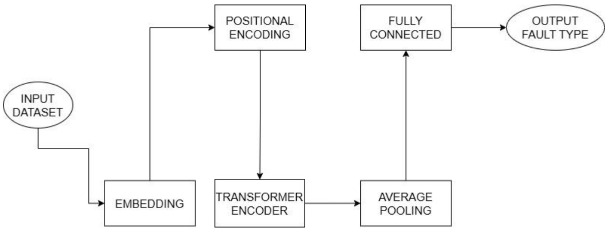

Figure 10: Flowchart of the Deep Learning Transformer model

The Transformer model is designed to process input features representing gas concentrations, which are first subjected to an embedding layer to map the input data into a higher-dimensional space. This process enhances the model's ability to capture the intricate relationships between different gas concentrations. The embedded inputs are then supplemented with positional encoding, which is crucial for retaining the order and structure of the input sequence, despite the inherent lack of sequential information in the input data. Figure 10 how the flowchart of how the deep learning model works.

The core of our model is the Transformer encoder, which leverages multi-head self-attention mechanisms and feedforward neural networks to extract deep, context-aware features from the input data. The use of multiple encoder layers allows the model to learn hierarchical representations of the input, which is essential for accurate fault classification.

To further refine the extracted features, an average pooling operation is applied, reducing the dimensionality and focusing on the most relevant information. The pooled features are then passed through a fully connected layer, which serves as the final classifier, outputting the predicted fault type.

To ensure the robustness of our model, we employed several data preprocessing steps, including the use of Synthetic Minority Over-sampling Technique (SMOTE) to address class imbalance and the removal of missing values to improve data quality. The dataset was then split into training and testing sets, with the training data used to optimize the model's parameters through backpropagation. The model was trained for 200 epochs, with performance monitored through accuracy metrics. The results demonstrate that the Transformer model is capable of achieving significant improvements in fault classification accuracy, surpassing traditional machine learning approaches and bringing us closer to our goal of 80% or more accuracy.

The methodology described above was systematically applied to three different datasets of varying sizes to evaluate the scalability and robustness of the Transformer model. Each dataset was processed through multiple stages, including feature engineering and synthetic data generation using SMOTE, to assess the model's performance across diverse data configurations. The following table 2 summarizes the number of rows and columns for each dataset across the different stages:

Table 2: Dataset sizes

Dataset size	Stage	Dimensions
Small	Original	(70, 7)
	Feature Engineering	(66, 14)
	SMOTE + Feature Engineering	(136, 14)
Medium	Original	(114, 8)
	Feature Engineering	(114, 13)
	SMOTE + Feature Engineering	(270, 13)
Large	Original	(231, 6)
	Feature Engineering	(213, 9)
	SMOTE + Feature Engineering	(465, 9)

In the small dataset, the original data consisted of 70 rows and 7 columns, representing the concentrations of dissolved gases. After applying feature engineering techniques, the dataset was expanded to 66 rows and 14 columns, where additional features were derived to capture more complex patterns. Applying SMOTE to this engineered dataset resulted in a larger dataset of 136 rows and 14 columns, addressing the class imbalance issue and providing a more comprehensive training set for the Transformer model.

The medium-sized dataset began with 114 rows and 8 columns. Feature engineering increased the dimensionality to 13 columns while maintaining the same number of rows. After applying SMOTE, the dataset expanded to 270 rows, further enriching the training data. Similarly, the large dataset, which initially had 231 rows and 6 columns, was transformed through feature engineering into a dataset with 213 rows and 9 columns. SMOTE application resulted in an expanded dataset with 465 rows and 9 columns, providing ample data for model training.

By following this methodology across datasets of varying sizes, we were able to demonstrate the Transformer model's adaptability and consistency in handling different data volumes. The results from these experiments validate the model's potential for generalizing across different datasets, making it a robust tool for fault classification in power transformers. This approach also highlights the importance of data preprocessing and feature engineering in enhancing the performance of deep learning models.

4. IMPLEMENTATION OF TRANSFORMER MODEL FOR FAULT PREDICTION IN TRANSFORMERS

The Transformer model was evaluated on three datasets of varying sizes, each subjected to different stages of data processing: original, feature-engineered, and SMOTE-enhanced feature engineering.

The following algorithm delineates the systematic approach employed for the classification and prediction of fault types in power transformers utilizing a Transformer-based deep learning framework.

4.1 Algorithm:

Step 1: Data Preparation

- Three datasets—categorized as small, medium, and large—were curated, as elaborated in Section 2.1 (Data Overview and Preprocessing).
- These datasets were selected to assess the scalability, robustness, and efficacy of the proposed Transformer model.

Step 2: Data Preprocessing

- The raw datasets were subjected to preprocessing to eliminate missing values (NAs) and to incorporate additional derived features through feature engineering, as detailed in Sections 2.2 and 2.3.
- Feature engineering was undertaken to enhance the representational capacity of the input data and to facilitate the extraction of meaningful patterns for fault classification.

Step 3: Data Partitioning

- The preprocessed datasets were partitioned into training (80%) and testing (20%) subsets to facilitate model training and performance evaluation.
- This stratified division ensures the reliability and generalizability of the proposed methodology.

Step 4: Data Standardization

- Standardization was applied post-class imbalance resolution (via SMOTE) and feature engineering to normalize the datasets.
- Each feature was transformed to have a mean of 0 and a standard deviation of 1. This ensures the mitigation of scale disparity across features, accelerates model convergence, and enhances predictive performance.

Step 5: Model Architecture and Training

The standardized data was fed into a Transformer-based architecture comprising the following sequential components:

i. **Embedding Layer:** Encodes input features into dense vector representations to facilitate model comprehension.

ii. **Positional Encoding:** Introduces positional context into the embeddings, ensuring the model captures the inherent ordering of features.

iii. **Transformer Encoder:** Leverages self-attention mechanisms and feed-forward networks to model intricate dependencies and relationships within high-dimensional data.

iv. **Average Pooling Layer:** Aggregates feature representations to create a compact latent space representation.

v. **Fully Connected Layer:** Maps the latent representation to a high-level feature space for fault classification.

vi. **Output Layer:** Outputs a probability distribution over fault types, with dimensionality corresponding to the number of fault categories.

Step 6: Model Evaluation and Performance Comparison

- The performance of the Transformer-based model was assessed across three dataset variants:
  - Original dataset.
  - Feature-engineered dataset.
  - SMOTE-enhanced dataset.
- Accuracy metrics were computed for each dataset variant, and comparative analysis was conducted to evaluate the impact of feature engineering and data augmentation (via SMOTE) on classification efficacy.

This implementation underscores the viability of Transformer-based deep learning architectures in addressing the challenges of high-dimensional and imbalanced datasets for fault diagnosis in power transformers. The proposed methodology advances the state-of-the-art in fault classification by leveraging feature engineering, SMOTE, and self-attention mechanisms to achieve superior predictive accuracy.

The results, summarized in Table 3, reveal significant variations in accuracy across the different datasets and preprocessing stages, reflecting the impact of data preparation and augmentation on model performance.

For the small dataset, the Transformer model achieved perfect accuracy (100%) on both the original and feature-engineered versions, indicating that the model was able to learn the relationships within the gas concentrations effectively without requiring additional synthetic data. However, when SMOTE was applied to address class imbalance, the accuracy dropped to 71.43%. This decrease suggests that while SMOTE successfully increased the dataset's size, it may have introduced noise or less representative samples that hindered the model's performance.

In the medium dataset, the original dataset yielded a moderate accuracy of 60.87%, which improved slightly to 65.22% after feature engineering. This improvement highlights the benefits of generating additional features to capture more complex relationships in the data. The most significant gain was observed when SMOTE was applied, with the accuracy jumping to 88.89%. This substantial improvement demonstrates the effectiveness of SMOTE in enhancing the training set's representativeness, allowing the Transformer model to generalize better to unseen data.

Table 3: Accuracies after Implementation of Transformer model for Fault Prediction in Transformers

Dataset Size	Stage	%Accuracy
Small	Original	100
	Feature Engineering	100
	SMOTE + Feature Engineering	71.43
Medium	Original	60.87
	Feature Engineering	65.22
	SMOTE + Feature Engineering	88.89
Large	Original	61.7
	Feature Engineering	74.42
	SMOTE + Feature Engineering	89.25

The large dataset exhibited similar trends, with the original dataset yielding an accuracy of 61.7%, which increased to 74.42% after feature engineering. This result underscores the importance of feature engineering in improving model performance, particularly when dealing with larger datasets. The application of SMOTE further boosted the accuracy to 89.25%, highlighting the importance of addressing class imbalance in large datasets. The hyperparameter adjustments made for the SMOTE-enhanced dataset, particularly the reduction in the number of heads and layers, likely contributed to the model's ability to handle the more complex and diverse training data effectively.

The choice of hyperparameters played a crucial role in the performance of the Transformer model across the different datasets and processing stages. For the small and medium datasets, consistent hyperparameters were applied, including a hidden dimension of 64, feed-forward dimension of 128, four attention heads, four layers, and a dropout rate of 0.1. The input dimension and number of classes were adjusted according to the dataset's specific characteristics, with the input dimension ranging from 6 to 13 and the number of classes from 4 to 6. The large dataset required more careful tuning, particularly after applying SMOTE. For the original large dataset, four attention heads and four layers were maintained, but for the feature-engineered and SMOTE-enhanced datasets, the number of heads was reduced to 2, and the number of layers to 3, reflecting the need for a more streamlined architecture to handle the increased data complexity. Across all datasets, a learning rate of 0.001 and 200 training epochs were used, ensuring sufficient training time for convergence without overfitting.

Overall, these results underscore the importance of data preprocessing and augmentation in enhancing the performance of deep learning models. The consistent improvements observed after applying feature engineering and SMOTE across all dataset sizes validate the robustness and adaptability of the Transformer model in classifying power transformer fault types. The model's performance on the large SMOTE-enhanced dataset indicates its potential for deployment in real-world scenarios where data diversity and class imbalance are common challenges.

5. CONCLUSION

This study demonstrates the effectiveness of using a Transformer-based model for classifying power transformer fault types based on gas concentration levels. By evaluating the model across three datasets of varying sizes and processing stages—original, feature-engineered, and SMOTE-enhanced—it was evident that both feature engineering and data augmentation significantly contributed to improved model accuracy. The model performed exceptionally well on the small dataset, achieving 100% accuracy in the original and feature-engineered stages, though the accuracy dropped to 71.43% after applying SMOTE. The medium and large datasets also showed substantial improvements with the application of SMOTE, with accuracies of 88.89% and 89.25%, respectively, indicating the model's potential to generalize well across diverse and imbalanced data.

The results underscore the importance of comprehensive data preprocessing, careful hyperparameter tuning, and the use of advanced deep learning techniques like Transformers in tackling complex classification tasks in the power systems domain. The hyperparameter adjustments made for the large dataset, particularly in reducing the number of attention heads and layers after applying SMOTE, highlight the necessity of optimizing model architecture to handle increased data complexity effectively.

However, the reliance on synthetic data generated through SMOTE raises concerns about the model's real-world applicability. While SMOTE helps to balance the dataset and improve model performance, it can introduce synthetic patterns that do not entirely represent real-world scenarios. Therefore, future work should focus on gathering more real-time data that captures various gas concentration levels and associated fault types in power transformers. This would enable a more robust evaluation of the Transformer model's performance in practical settings and ensure that the predictions are not overly influenced by synthetic data patterns. Expanding the dataset with real-world measurements will provide a more reliable basis for deploying this model in operational environments, ultimately enhancing its utility in preventing power transformer failures.

6. ACKNOWLEDGMENTS

The authors wish to thank the University of West Florida for providing the funds and facilities to conduct this research.

7. REFERENCES

[1] Duval, M. (2002). A review of faults detectable by gas-in-oil analysis in transformers. _IEEE Electrical Insulation Magazine_, 18(3), 8-17.

[2] Vaswani, A., et al. (2017). Attention is All You Need. In _Proceedings of the 31st International Conference on Neural Information Processing Systems_, 5998–6008.

[3] Li, Z., et al. (2023). Research on the transformer fault diagnosis method based on LSTM artificial neural network and DGA. _2022 International Conference on Intelligent Computing and Machine Learning (2ICML)_, IEEE, 2023.

[4] Zhang, Y., et al. (2022). Fault diagnosis of transformer using artificial intelligence: A review. _Frontiers in Energy Research_, 10, 1006474.

[5] Chawla, N. V., et al. (2002). SMOTE: Synthetic Minority Over-sampling Technique. _Journal of Artificial Intelligence Research_, 16, 321–357.

[6] He, H., & Garcia, E. A. (2009). Learning from Imbalanced Data. _IEEE Transactions on Knowledge and Data Engineering_, 21(9), 1263-1284.

[7] Wang, S., & Yao, X. (2012). Multiclass Imbalance Problems: Analysis and Potential Solutions. _IEEE Transactions on Systems, Man, and Cybernetics_, 42(4), 1119-1130.

[8] Y.T. Yu, M.F. Lau. (2005). A comparison of MC/DC, MUMCUT and several other coverage criteria for logical decisions, Journal of Systems and Software, 2005, in press.

[9] Illias, H.A., Chai, X.R., Abu Bakar, A.H. and Mokhlis, H., 2015. Transformer incipient fault prediction using combined artificial neural network and various particle swarm optimisation techniques. PloS one, 10(6), p.e0129363.

[10] Duval, M. and DePabla, A., 2001. Interpretation of gas-in-oil analysis using new IEC publication 60599 and IEC TC 10 databases. IEEE Electrical Insulation Magazine, 17(2), pp.31-41.

[11] Seifeddine, S., Khmais, B. and Abdelkader, C., 2012, March. Power transformer fault diagnosis based on dissolved gas analysis by artificial neural network. In 2012 first international conference on renewable energies and vehicular technology (pp. 230-236). IEEE.

[12] N. V. Chawla et al., (2002). SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research, vol. 16, pp. 321-357.

[13] G. Lemaître, L. Nogueira, and J. L. D. Carvalho,. (2017). SMOTE: Synthetic Minority Over-sampling Technique. Journal of Machine Learning Research, vol. 18, pp. 1-23.

[14] V. K. Elaparthy, N. R. S. T. S. Suresh, and S. M. Sharmila, (2019). A hybrid SMOTE model for detecting fraudulent transactions in imbalanced datasets. Journal of King Saud University-Computer and Information Sciences.

[15] M. Ahmed, M. Ganaie, and R. Bhat, (2020). An improved hybrid model for transformer fault detection and prediction using SMOTE. Journal of Electrical Engineering & Technology, vol. 15, no. 2, pp. 625-635.

[16] Devlin, J., (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

[17] A. Radford et al., (2018). Improving Language Understanding by Generative Pre-Training. OpenAI.

W. Zhang et al., (2020). Transformer-based approach for fault diagnosis in power systems. IEEE Transactions on Industrial Electronics, vol. 67, no. 8, pp. 6718-6727.