Paper ID – 2433
Asset Management for Power Systems

Advancing Transformer Diagnostics: A Statistical Analysis of a Publicly Available DGA Database

Mohamed Rayan BARHDADI
Texas A&M University
Doha, Qatar
rayan.barhdadi@tamu.edu

Farheen F. JALDURGAM
Texas A&M University
Doha, Qatar
farheen.jaldurgam@qatar.tamu.edu

Selma AWADALLAH
Hamad Bin Khalifa University
Doha, Qatar
sawadallah@hbku.edu.qa

SUMMARY

This paper presents the development and statistical analysis of a comprehensive, publicly accessible DGA database designed to address this critical research gap. The database encompasses 743 transformer records following rigorous data acquisition procedures based on standard preprocessing methodologies. The dataset incorporates measurements of five key hydrocarbon gases: hydrogen (H_2), methane (CH_4), ethylene (C_2H_4), ethane (C_2H_6), and acetylene (C_2H_2), collected from diverse geographical locations and transformer types. A systematic data quality verification methodology was implemented, utilizing format normalization and duplicate removal to ensure data integrity while preserving all diagnostic information, including fault signatures. The final dataset contains 743 transformers with complete gas measurements, with some maintaining diagnostic utility with fault labels across multiple categories. Statistical analysis of 743 transformer records showed mean concentrations of 348.42 ppm (H_2), 282.02 ppm (C_2H_4), and 160.21 ppm (CH_4). All gases exhibited high positive skewness (7.76-16.94) with concentration ranges spanning four orders of magnitude. Correlation analysis revealed strong relationships among hydrocarbon gases (0.65-0.85), with the highest correlation between CH_4 and C_2H_6 (0.846). The research contributes a web-based platform (https://bmrayan.github.io/dgadb/) that provides intuitive access to the database through structured query capabilities, statistical analysis with data visualization, and comprehensive documentation of data collection methods. The platform features an interactive query interface with predefined quick queries for common diagnostic scenarios, enabling researchers and practitioners to efficiently extract data based on specific fault conditions or gas concentration thresholds. This standardized resource enables benchmarking of diagnostic algorithms, supports machine learning model development, and facilitates collaborative research in transformer condition assessment, ultimately contributing to enhanced grid reliability and optimized maintenance strategies.

KEYWORDS

Dissolved Gas Analysis; Power Transformer; Database; Statistical Analysis; Web Interface

1. INTRODUCTION

Power transformers are the backbone of modern transmission and distribution grids, and their dependable service is central to overall network stability [1]. Over the years of operation, these high-value units experience thermal, electrical, and mechanical stresses that weaken insulation and can ultimately lead to failure [2]. Since the late 1960s, Dissolved Gas Analysis (DGA) has been the principal non-intrusive technique for detecting such incipient faults [3,4], and gas thresholds and ratio rules are set out in IEEE Std C57.104-2008 [5]. Industry surveys still list unexpected transformer outages among the costliest reliability events for utilities [1]. Further improving diagnostic accuracy requires broad, standardized data, yet very few large public DGA datasets exist. Most utilities keep their records private, and the few datasets that are available differ in reporting units, data-cleaning steps, and fault labels [6,7]. Early efforts to compile DGA records from multiple sources encountered incompatible measurement units and ad-hoc filtering procedures that hindered comparison with other studies [8]. The few public DGA repositories released to date contain fewer than a few hundred transformers, limiting statistical robustness.

Researchers have explored statistical approaches to set more representative gas levels. Zhao et al. [9] fitted log-normal and Weibull distributions to 760 Chinese transformers and derived "typical" and "alarm" values that aligned better with field experience than standardized tables. IEC 60599:2015 likewise notes that diagnosis improves when data are cleaned and interpreted uniformly across regions and equipment designs [10]. These findings motivate the creation of larger, carefully curated datasets to strengthen DGA monitoring across diverse transformer fleets.

To address these challenges, this paper introduces a 743-transformer database with systematic data cleaning and quality control, delivered through an open web interface offering structured queries and comprehensive visualization. To provide a comprehensive understanding of the distribution and variability of DGA gas concentration, statistical analysis has been performed using histograms and boxplots for data visualization, along with the calculation of parameters such as mean, standard deviation, maximum, skewness, and kurtosis. The dataset aims to offer researchers and practitioners a common, transparent baseline for future studies and diagnostic algorithm development.

2. DATA COLLECTION AND ORGANIZATION

2.1 Data Acquisition Methodology

The database compilation process involved systematic collection from multiple geographic regions and sources spanning peer-reviewed scientific literature, industry technical reports, and publicly available monitoring data. Sources were selected based on data quality criteria, including measurement methodology documentation, information availability, and adherence to standard gas measurement units (parts per million by volume). The scope encompasses transformers with both dated and undated measurements, providing comprehensive coverage of evolving transformer designs, insulation systems, and operating conditions. Geographic diversity was ensured through the inclusion of data from North America, Europe, South Asia, and other regions, representing various climatic conditions, loading patterns, and maintenance practices.

2.2 Database Structure and Organization

The database architecture employs a relational structure implemented in SQL (Structured Query Language) to facilitate efficient data management and query operations. The core entity-relationship model consists of:

a) Transformer Table: Primary entity containing unique transformer identifiers, technical specifications where available, and operational metadata
b) Measurement Table: Gas concentration records linked to transformers through foreign key relationships
c) Fault Classification Table: Standardized fault categories based on IEEE and IEC terminology [5,10].

Each record contains measurements for five key hydrocarbon gases recognized as primary diagnostic indicators:

a) Hydrogen (H2): Indicator of partial discharge and corona activity
b) Methane (CH4): Low-temperature thermal decomposition marker
c) Ethylene (C2H4): Medium to high-temperature thermal fault indicator
d) Ethane (C2H6): Low-temperature thermal fault marker
e) Acetylene (C2H2): High-temperature thermal fault and arcing indicator

2.3 Data Cleansing Methodology

Initial data preprocessing addressed common quality issues, including:
a) Format Normalization: Standardization of numerical representations and delimiter conventions
b) Duplicate Detection: Identification and elimination of redundant entries through hash-based comparison

The cleaning process maintained the complete dataset of 743 transformers with full gas concentration measurements, ensuring preservation of all fault signatures and diagnostic patterns essential for algorithm validation and development. Figure 1 shows the scatter plots of the distribution of five dissolved gases (H2, CH4, C2H4, C2H6, C2H2) across all transformer samples.

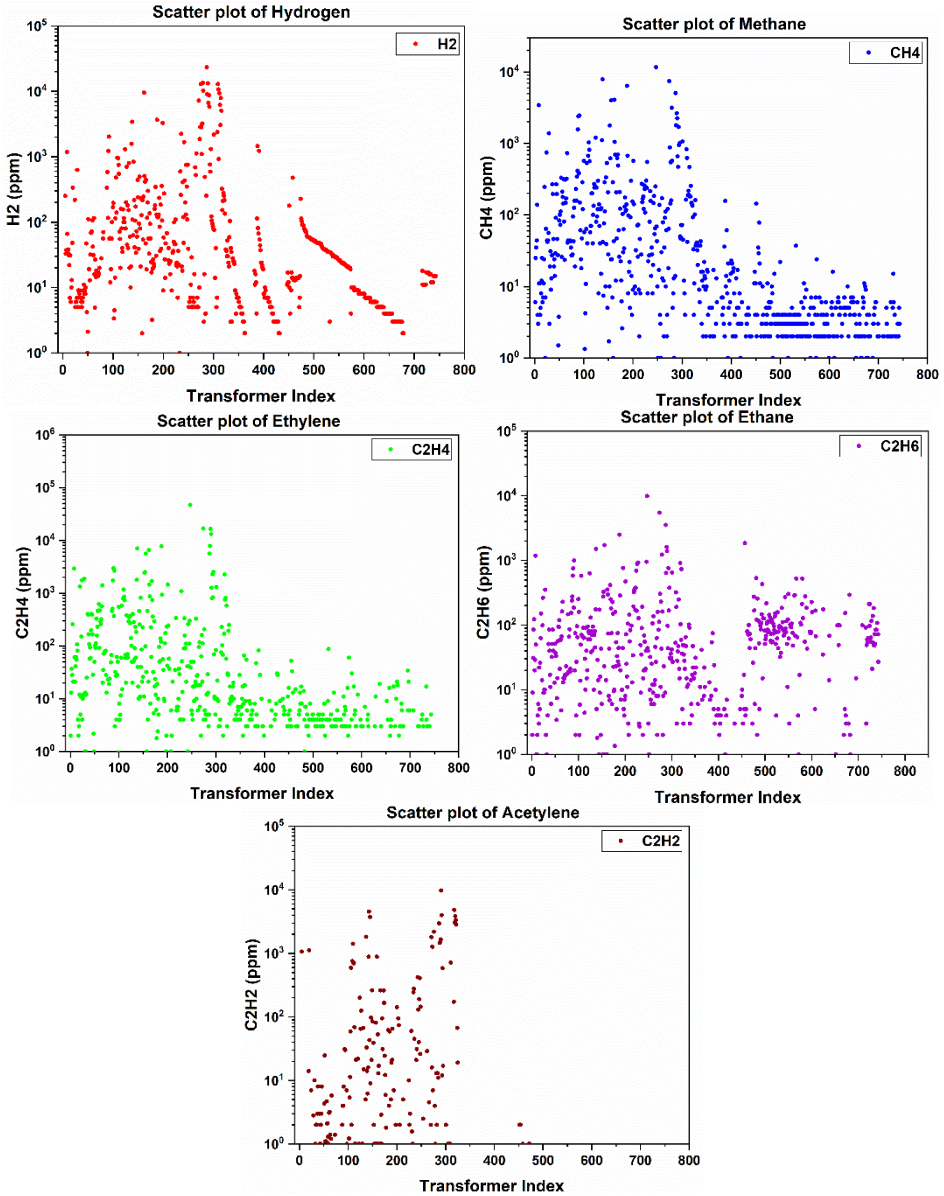

Figure 1. Scatter plots show the distribution of five dissolved gases (H2, CH4, C2H4, C2H6, C2H2) across all transformer samples. Each subplot displays concentration values on a log scale to accommodate the wide range of measurements.

3. WEB INTERFACE DEVELOPMENT

3.1 Platform Architecture

The web-based access platform (https://bmrayan.github.io/dgadb/) was developed using modern web technologies to provide intuitive access to the database without requiring specialized database management expertise. The interface of the webpage is shown in Figure 2.

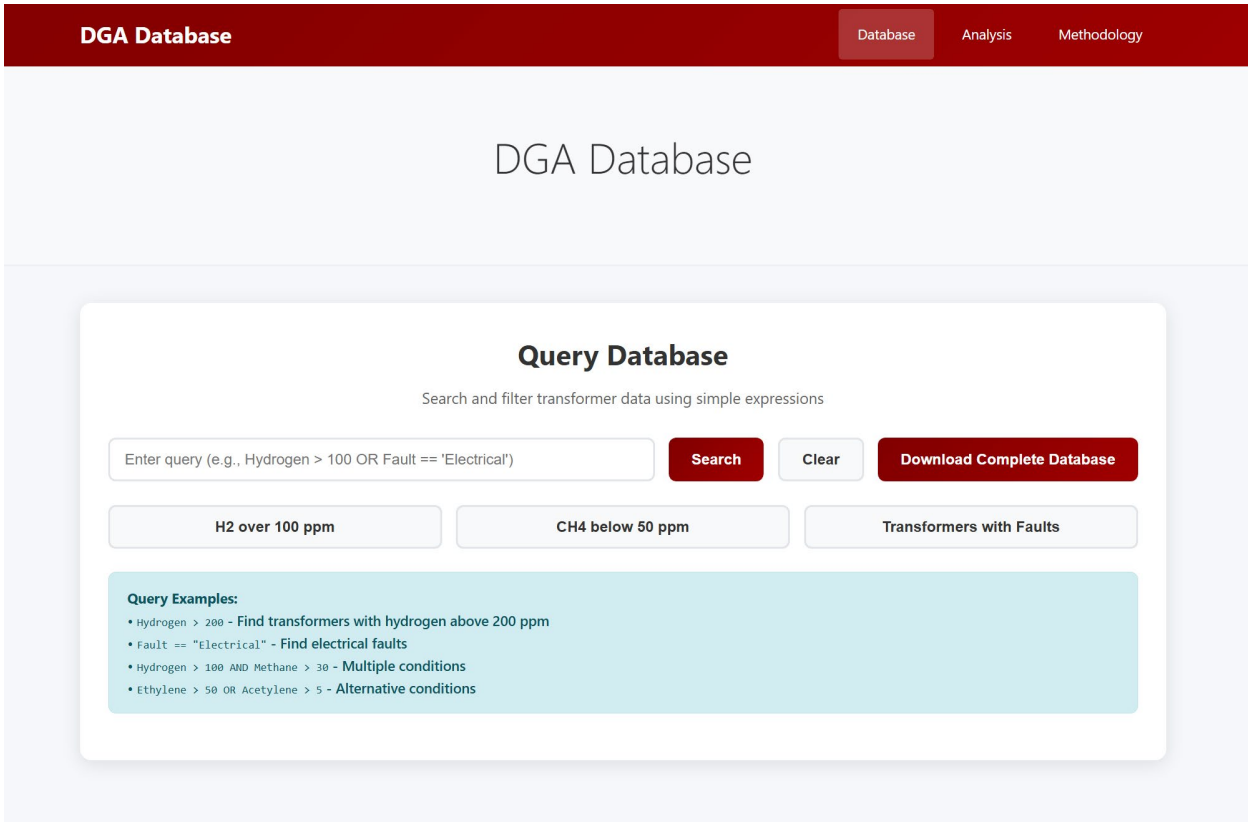

Figure 2. The webpage interface shows the main interface, which is organized into three functional sections: database, statistical analysis, and data-collection method.

3.1.1 Database Query Interface

The query interface provides users with flexible data extraction capabilities through:

a) Parameter-based filtering by gas, concentration range, or fault category
b) Advanced query builder for multi-criteria searches
c) Quick-query buttons for common scenarios (“High Hydrogen,” “High Methane,” etc.)
d) SQL-like architecture
e) Export options in CSV, JSON, or XML

The interface uses a clean layout with a red navigation bar and a placeholder prompt (“Type your query, e.g., Hydrogen > 50 ppm”) to guide users.

3.1.2 Statistical Analysis Module

Integrated statistical analysis capabilities include:

a) Descriptive Statistics: Real-time calculation of mean, median, standard deviation, skewness, and kurtosis
b) Correlation Analysis: Inter-gas relationship quantification with statistical significance testing
c) Distribution Analysis: Histogram generation and normality testing for the selected data subset

d) Fault Classification Analysis: Pie chart depicting the distribution of fault types in the dataset

3.2 User Experience Design

Responsive design ensures full functionality on desktops, tablets, and mobile devices. User help is provided through in-page documentation and example queries.

4. STATISTICAL ANALYSIS RESULTS

4.1 Descriptive Statistics

The complete dataset of 743 transformers provides comprehensive statistical characteristics across all five diagnostic gases. The database maintains a full representation of transformer operating conditions and fault signatures without artificial filtering constraints. The summary of the descriptive statistics of the DGA gas data is given in Table 1. The statistical analysis reveals significant variation in gas concentrations, with hydrogen showing the highest mean concentration at 348.42 ppm, followed by ethylene at 282.02 ppm. Methane and acetylene demonstrate intermediate concentrations at 160.21 ppm and 93.69 ppm, respectively, while ethane shows the lowest mean at 107.17 ppm. Distribution characteristics indicate positive skewness across all gases, with ethylene exhibiting the highest skewness (16.94) and hydrogen the lowest (7.76). Kurtosis values are also elevated, ranging from 73.37 for hydrogen to 348.86 for ethylene, indicating heavy-tailed distributions that are characteristic of DGA data. The box plots and histograms of the DGA gases are given in Figure 3 and Figure 4, respectively.

Table 1. Descriptive Statistics of the different gases of the DGA data.

Gas (ppm)	Count	Mean	Std Dev	Max	Skewness	Kurtosis
Hydrogen (H2)	743	348.4162	1674.116	23349	7.762784	73.36693
Methane (CH4)	743	160.2106	746.236	11646	9.627752	113.3367
Ethylene (C2H4)	743	282.0246	2089.029	46976	16.94388	348.8592
Ethane (C2H6)	743	107.1718	479.3185	9901	14.39497	260.3331
Acetylene (C2H2)	742	93.68542	573.4672	9740	9.979591	129.3289

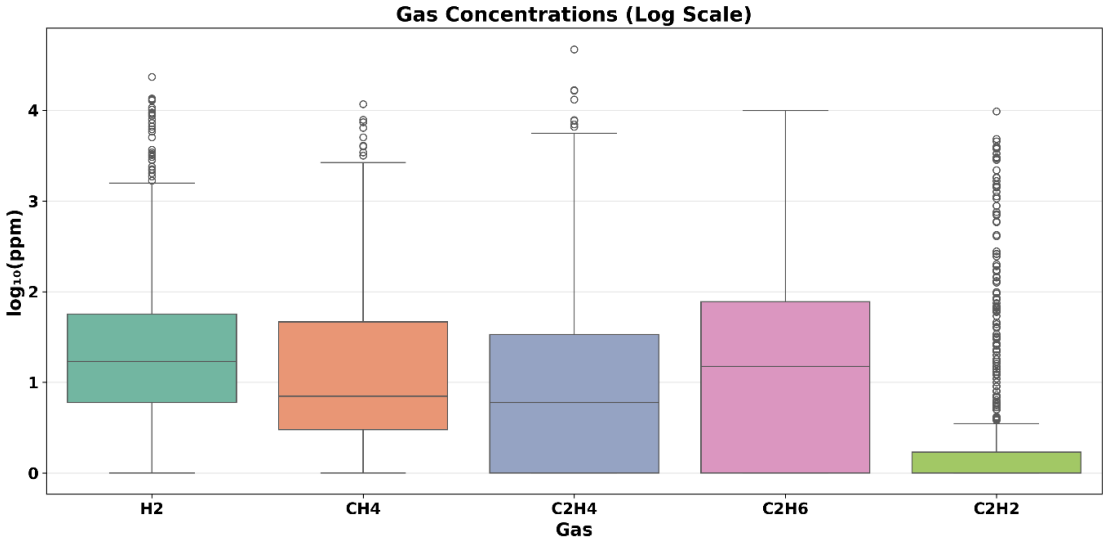

Figure 3. Box plots of dissolved gas concentrations displayed on a logarithmic scale, showing median values, quartiles, and outliers for each gas type across the transformer dataset.

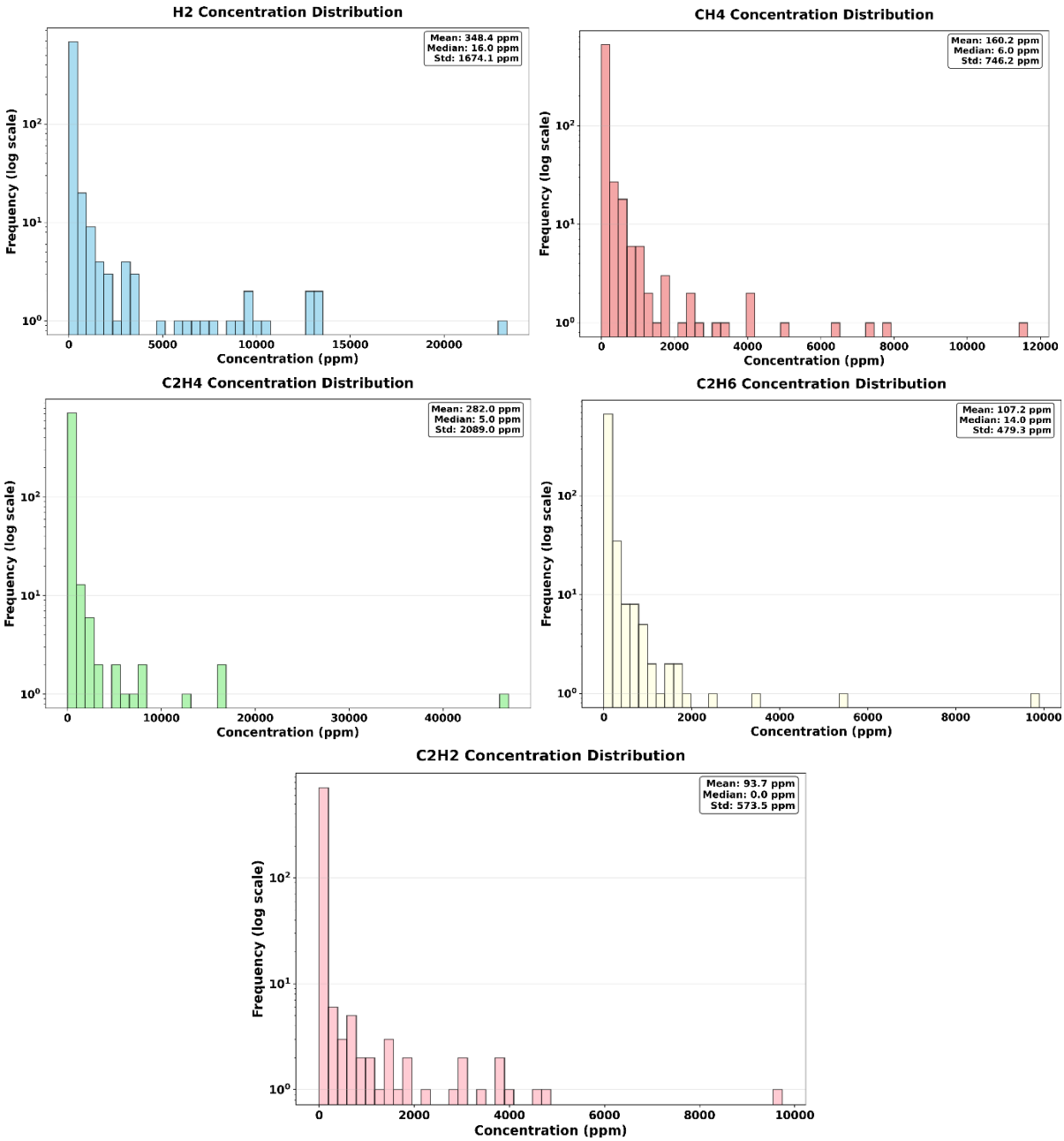

Figure 4. Histograms showing the frequency distribution of each dissolved gas concentration with logarithmic y-axis scaling to highlight the distribution characteristics across different concentration ranges.

4.2 Correlation Analysis

Inter-gas correlations demonstrate expected relationships based on fault generation mechanisms. Figure 5 shows the correlation heatmap showing the pairwise Pearson correlation coefficients between five gases. Methane and ethylene show a strong positive correlation (0.81), consistent with thermal fault progression patterns. Ethylene and ethane exhibit the highest correlation (0.90), reflecting their common thermal decomposition origins. Hydrogen shows moderate correlations with other gases, ranging from 0.27 to 0.41, indicating its association with multiple fault types.

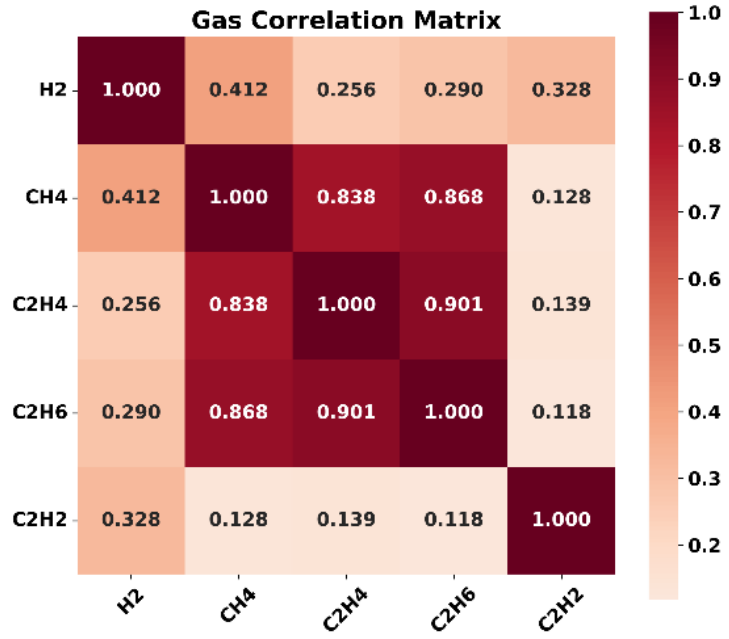

Figure 5. Inter-Gas Correlation Matrix for DGA Parameters, Pearson correlation coefficients between dissolved gas concentrations. The heatmap reveals strong correlations among hydrocarbon gases and moderate correlations with hydrogen.

4.3 Distribution Analysis

Gas concentration distributions follow characteristic patterns expected in DGA data, with normal characteristics enabling appropriate statistical modeling approaches. The preservation of complete data ranges ensures representation of both normal operating conditions and fault signatures essential for diagnostic algorithm development. The normal distribution analysis of DGA gas concentrations is given in Figure 6.

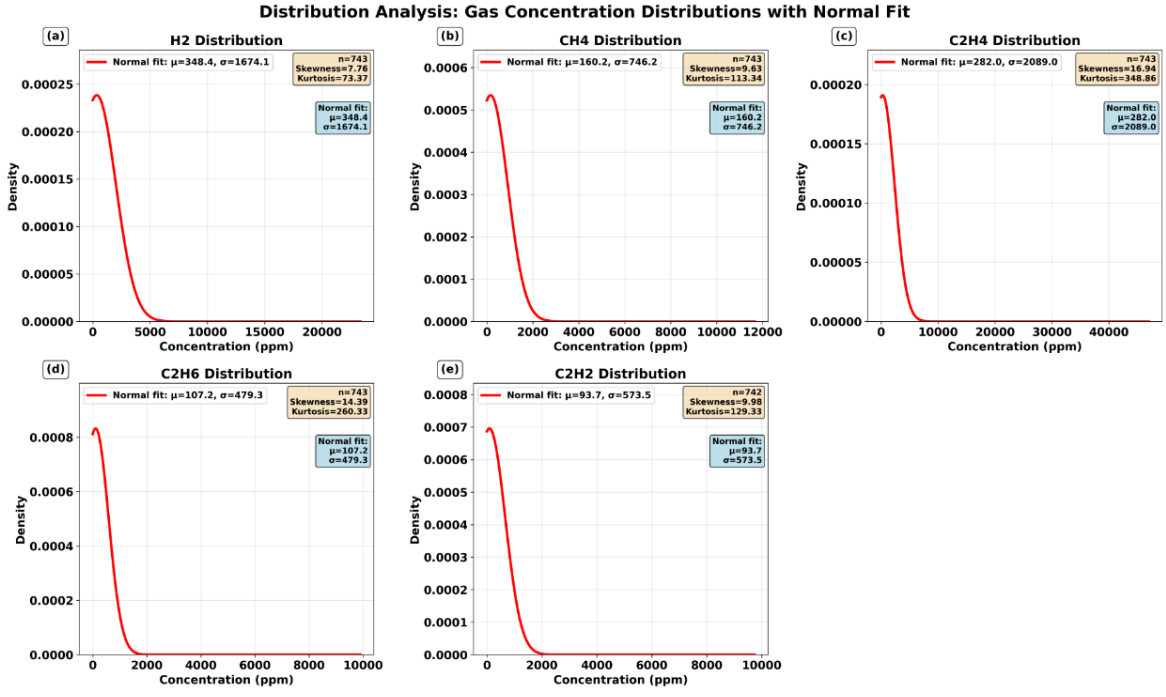

Figure 6. Normal distribution analysis of DGA gas concentrations. Normal distribution fitting for dissolved gas concentrations in actual ppm values. Each subplot shows the histogram of gas

concentration data with fitted normal distribution curves, including statistical parameters (μ, σ) and distribution characteristics, demonstrating the non-normal nature of raw DGA data and the need for specialized statistical approaches.

4.4 Fault Classification Analysis

The database contains 18 transformers with fault classifications, providing representation for diagnostic validation. While this represents a subset of the total dataset, these labeled cases offer valuable insights into fault-specific gas generation patterns essential for algorithm development and validation. Figure 7 depicts the pie chart and mean gas concentration distribution for different fault types.

The distribution of faults reveals:

a) No Fault: 14 cases (77.8%) - These transformers exhibit normal operating conditions, providing baseline gas concentration profiles against which fault conditions can be compared. The predominance of no-fault cases reflects typical field conditions where the majority of monitored transformers operate normally.
b) Electrical Fault: 2 cases (11.1%) - These cases demonstrate gas patterns associated with electrical discharge phenomena, typically characterized by elevated acetylene (C_2H_2) production and hydrogen (H_2) generation. Despite the limited sample size, these cases capture the distinctive gas signatures of electrical stress conditions.

c) Thermal Fault: 2 cases (11.1%) - These transformers exhibit gas patterns indicative of thermal degradation, generally showing increased production of ethylene (C_2H_4) and ethane (C_2H_6) as cellulose and oil decompose under elevated temperatures. This fault distribution represents authentic transformer operating conditions and diagnostic scenarios encountered in practice.

This fault distribution, while limited in absolute numbers, represents authentic transformer operating conditions and diagnostic scenarios encountered in practice. The analysis of mean gas concentrations across fault categories reveals distinct gas signature patterns, with electrical faults showing pronounced acetylene peaks and thermal faults demonstrating elevated hydrocarbon gas levels. These characteristic patterns, even with limited samples, align with established DGA interpretation frameworks and provide validation benchmarks for diagnostic algorithm development.

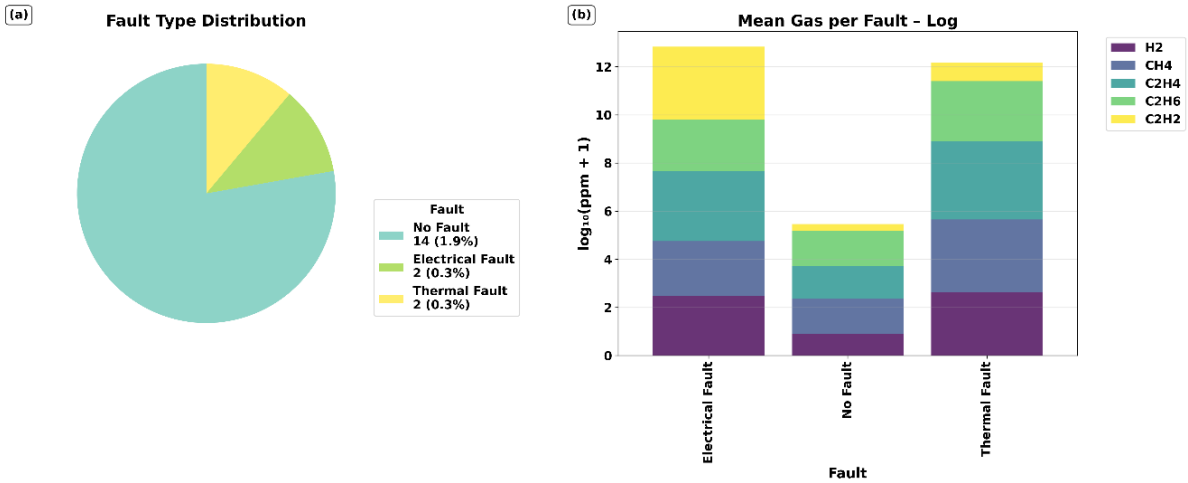

Figure 7. Fault Type Distribution and Gas Signature Analysis, (a) Pie chart showing the distribution of fault types in the dataset. (b) Mean gas concentrations by fault type are displayed on a logarithmic scale, revealing characteristic gas signatures for different transformer fault conditions.

5. APPLICATIONS, LIMITATIONS, AND FUTURE DIRECTIONS

5.1 Research Impact

The dataset of 743 transformer records enables practical applications across multiple domains. In machine learning development, researchers can train and validate classification algorithms using standardized data formats, ensuring consistent model comparison and reproducible results. The dataset supports direct benchmarking of established diagnostic methods, including IEEE C57.104 thresholds and Duval Triangle interpretation, allowing systematic evaluation of diagnostic accuracy across different approaches. Educational institutions benefit from access to real-world DGA data for teaching interpretation techniques, statistical analysis, and student project development. Industry practitioners utilize the dataset as a standardized benchmark for validating proprietary diagnostic systems and assessing commercial DGA software performance. The 18 fault-labeled cases, while limited, provide essential validation examples for supervised learning approaches and classification accuracy assessment.

5.2 Limitations and Future Improvements

Several limitations affect the dataset's current scope:

a) Data Completeness: Absence of transformer specifications (power rating, voltage class, manufacturing year) and measurement uncertainty documentation for historical records
b) Geographic Coverage: Overrepresentation of certain regions, limiting generalizability across different climates and operational practices
c) Fault Representation: Fault Representation: Only 18 fault-labeled cases with 77.8% being no-fault conditions, requiring resampling for balanced classifier development
d) Temporal Gaps: Most transformers lack dating information, preventing age-related degradation studies

Future development priorities include:

a) Data Expansion: Annual addition of 200-300 transformers with complete metadata, emphasizing fault cases and geographic diversity
b) Quality Enhancement: Automated outlier detection, cross-validation with laboratory standards, and community-driven verification processes
c) Metadata Integration: Systematic incorporation of transformer specifications, operational history, and environmental conditions, where available

6. CODE AVAILABILITY

To ensure reproducibility and facilitate further research, all computational resources developed for this study are publicly available through a comprehensive GitHub repository. The complete codebase includes data processing algorithms, visualization tools, web interface implementation, and comprehensive documentation under the MIT open-source license.

Repository Access: [GitHub Repository URL - https://github.com/bmrayan/DGA-Dataset]

7. CONCLUSION

This research addresses the critical challenge of limited public access to high-quality transformer dissolved gas analysis (DGA) data by developing a comprehensive database of 743 transformer records. The dataset represents a significant advancement in publicly available DGA resources, providing standardized, quality-controlled data essential for developing and validating diagnostic methodologies. Through systematic data collection and rigorous preprocessing, the database preserves complete diagnostic information while ensuring consistency across diverse data sources. Statistical analysis revealed all five monitored gases exhibit high positive skewness (7.76-16.94) with concentration ranges

spanning four orders of magnitude, establishing quantitative benchmarks for transformer condition assessment. Strong correlations among hydrocarbon gases (0.65-0.85) confirm established gas generation relationships, while distinct patterns in hydrogen and acetylene provide diagnostic differentiation capabilities essential for fault classification.

The web-based platform (https://bmrayan.github.io/dgadb/) transforms data accessibility through intuitive query interfaces, figures, and comprehensive documentation, democratizing DGA research for utilities, academic institutions, and researchers worldwide. This open-access approach breaks down traditional barriers where high-quality DGA data remained confined to proprietary systems. Despite limitations including incomplete transformer metadata and geographic bias, the dataset provides an immediate starting point for machine learning development, educational applications, and industrial benchmarking. The 18 fault-labeled cases offer crucial validation examples representing authentic field conditions with verified electrical and thermal fault signatures.

By establishing this foundational dataset and demonstrating its analytical potential, this research contributes to improving power transformer reliability through data-driven condition assessment. The commitment to maintaining this as a community-driven repository, with planned expansions and technical enhancements including API development, ensures its continued relevance as DGA practices advance. Public availability of both data and analysis tools represents a significant step toward collaborative advancement in transformer diagnostics, ultimately supporting more reliable and efficient power grid operations worldwide through evidence-based maintenance strategies.

BIBLIOGRAPHY

[1] CIGRE WG A2.37, "Transformer reliability survey," CIGRE Technical Brochure 642, 2015.
[2] IEEE Guide for Loading Mineral-Oil-Immersed Transformers and Step-Voltage Regulators, IEEE Std C57.91-2011, 2012.
[3] R. R. Rogers, "IEEE and IEC codes to interpret incipient faults in transformers using gas-in-oil analysis," IEEE Trans. Electrical Insulation, vol. EI-13, no. 5, Oct. 1978, pp. 349–354.
[4] M. Duval, "A review of faults detectable by gas-in-oil analysis in transformers," IEEE Electrical Insulation Magazine, vol. 18, no. 3, May/Jun. 2002, pp. 8–17.
[5] IEEE Guide for the Interpretation of Gases Generated in Oil-Immersed Transformers, IEEE Std C57.104-2008, 2008.
[6] T. Herath et al., "Data analytics for transformer dissolved-gas analysis to aid asset management," Proc. CIGRE Paris Session, Paper A2-10403, 2024.
[7] P. Mirowski and Y. LeCun, "Statistical machine learning and dissolved-gas analysis: A review," IEEE Trans. Power Delivery, vol. 27, no. 4, Oct. 2012, pp. 1791–1799.
[8] S. A. M. Abdelwahab et al., "Transformer fault diagnosis intelligent system based on DGA methods," Scientific Reports, vol. 14, article 27645, 2024.
[9] C. Zhao, Q. Huang, D. Li, H. Bai, and Y. Cheng, "The statistical distribution of the DGA data of transformers and its application," Proc. IEEE Int. Conf. High Voltage Engineering and Application, 2016, pp. 1–6.
[10] IEC 60599:2015, Mineral Oil-Filled Electrical Equipment in Service – Guidance on the Interpretation of Dissolved and Free Gas Analysis, IEC, 2015.
[11] IEEE Guide for the Interpretation of Gases Generated in Oil-Immersed Transformers, IEEE Std C57.104-2019, 2019.