International Journal of Inventive Engineering and Sciences (IJIES)
ISSN: 2319-9598 (Online), Volume-13 Issue-7, July 2026

Development of A Chromatographic Database for Applying Machine Learning for Drug Analysis

Ashwini M. Mawal*, Ishit S. Rindhe, Shaikh Hanif Ashfaque
Suresh K. Gaikwad, Mahesh Bhadane, Parag Chavan

Abstract: ***This study introduces a machine learning-based framework to optimise chromatographic data analysis in pharmaceutical quality assurance. We developed a chromatographic peak database and trained an XGBoost model—widely adopted in industry for its scalability and robustness across diverse data ranges—to predict the acceptability of chromatograms. The dataset included 100 CSV files representing real-world chromatographic waveforms, from which we extracted features such as retention time, peak height, and peak width for training. Our model achieved 65% accuracy, 3% precision, 38% recall, and an average RMSE, highlighting ML's potential to augment current manual quality assessment workflows. No benchmark model existed for this application; therefore, our comparative analysis with the Random Forest model provided the only available baseline. Unlike traditional, time-consuming manual reviews—where quality analysts may evaluate up to 100 chromatograms daily, each taking approximately five minutes—our predictive system streamlines the process by flagging only non-compliant cases, effectively reducing human workload by over 90%. This innovation lays foundational work for AI-driven chromatography, especially in areas where literature integrating ML with chromatographic analysis remains sparse.***

Keywords: ***Gas Chromatography, CSV, Data Processing, Waveform Plotting, Peak Identification, Training Data, AI Algorithm, Machine Learning, Model Building, Model Testing, Predicted Model, Manual Input***

Nomenclature:
HPLC: High-Performance Liquid Chromatography
AI: Artificial Intelligence
ML: Machine Learning
LC: Liquid Chromatography
GC: Gas Chromatography
FWHM: Full Width at Half Maximum
RMSE: Root Mean Square Error
QAs Quality Analysts

Manuscript received on 01 June 2026 | First Revised Manuscript received on 10 June 2026 | Second Revised Manuscript received on 02 July 2026 | Manuscript Accepted on 15 July 2026 | Manuscript published on 30 July 2026.

*Correspondence Author(s)
Ashwini M. Mawal*, Department of Chemistry, School of Science, Sandip University, Nashik, Maharashtra, India. Email ID: ashwinisrindhe@gmail.com
Ishit S. Rindhe, Department of Chemistry, School of Science, Sandip University, Nashik, Maharashtra, India. Email ID: hanif.drugchem2@gmail.com
Shaikh Hanif Ashfaque, Department of Chemistry, School of Science, Sandip University, Nashik, Maharashtra, India. Email ID: ishitSRindhe@gmail.com
Suresh K Gaikwad, Department of Chemistry, School of Science, Sandip University, Nashik, Maharashtra, India. Email ID: sureshgaikwad1990@gmail.com
Mahesh Bhadane, Department of Chemistry, School of Science, Sandip University, Nashik, Maharashtra, India. Email ID: maheshrbhadene@gmail.com
Dr. Parag Chavan, Assistant Professor, Department of Chemistry, School of Science, Sandip University, Nashik, Maharashtra, India. Email ID: parag247@gmail.com

© The Authors. Published by Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP). This is an open-access article under the CC-BY-NC-ND license http://creativecommons.org/licenses/by-nc-nd/4.0/

I. INTRODUCTION

Chromatographic techniques such as Gas Chromatography (GC) and High-Performance Liquid Chromatography (HPLC) are widely used in the pharmaceutical and chemical industries to separate, identify, and quantify compounds. These methods generate complex waveforms, or chromatograms, that require meticulous peak identification and interpretation—processes that remain predominantly manual in routine quality assurance workflows. In many industrial settings, Quality Analysts (QAs) must evaluate hundreds of chromatograms per day, with each review taking about 5 minutes. This cumulative workload creates significant challenges for time management, consistency, and accuracy, especially in high-throughput environments.

Despite advances in automation and analytics, a notable gap remains in applying artificial intelligence (AI) and machine learning (ML) to automate chromatographic interpretation. Recent ML studies have shown promise in spectral data analysis, sensor fusion, and process modelling; however, literature directly addressing AI-assisted peak detection and chromatogram validation remains limited. Most existing solutions still rely on rule-based algorithms or user-dependent preprocessing tools that lack predictive capabilities and scalability.

In this work, we propose a novel machine learning framework that integrates automated chromatographic peak analysis with intelligent model-driven decision-making. Our solution constructs a structured peak database derived from real chromatographic data. It employs XGBoost—a robust gradient boosting model known for its performance on tabular datasets—as the primary prediction engine. To validate this choice, we trained a comparative Random Forest model and compared the two algorithms' performance.

Our training dataset comprises 100 chromatographic files in CSV format and captures essential features such as retention time, peak height, and peak width. Preliminary results show that the XGBoost model achieved 65% accuracy, 3% precision, and 38% recall, with an average RMSE. While the current performance shows room for refinement, it also demonstrates the viability of ML-assisted workflows for chromatogram evaluation. Unlike traditional manual review, the proposed method proactively filters only potentially non-compliant chromatograms,

© Copyright: All rights reserved.

reducing analyst review time by over 90%.

This research addresses a critical operational bottleneck in pharmaceutical quality control and contributes to the emerging intersection of chromatographic science and machine learning. By laying a scalable data foundation and demonstrating practical ML integration, this work offers a forward-looking blueprint for intelligent, AI-enhanced analytical instrumentation.

Chromatography is a key lab technique for separating mixtures into their individual parts. It's especially useful in analytical chemistry for purifying and analysing different compounds. The process moves a mixture dissolved in a "mobile phase" through a "stationary phase," allowing components to separate based on how they interact with each phase [1]. Chromatography takes many forms, including gas chromatography (GC), liquid chromatography (LC), and high-performance liquid chromatography (HPLC), each suited to specific analyses [2].

II. PLOTTING CHROMATOGRAPHY DATA

Chromatographic data is often displayed as a chromatogram, which is a graph showing the detector response (y-axis) over time (x-axis). Each peak represents a different sample component. Plotting chromatography data involves visualising the output of a chromatography experiment, typically as a chromatogram. This is a graph showing detector response (e.g., absorbance, conductivity) versus time or volume [3].

Here's a step-by-step guide based on the flowchart provided:

Key aspects of data plotting include:

A. Receive Data

- Collect data from the chromatography system, often in CSV format, containing time and detector response [4].

B. Plot the Waveform:

- Use software (e.g., Excel, Python, R) to plot the data. The x-axis represents time or volume, and the y-axis represents detector response [5].

C. Identify Peaks:

- Analyse the chromatogram to identify peaks, which correspond to different compounds. Key parameters include retention time, peak height, and area [6].
  - i. *Peak Identification:* Determining which peaks correspond to which compounds.
  - ii. *Retention Time:* The time a compound takes to pass through the system, helping to identify it.
  - iii. *Peak Area/Height:* Used to measure the amount of each component.
  - iv. *Database:* Create a database of peaks vs. components.

D. Challenges in Chromatography

i. *Complex Data Analysis*

- **Challenge:** Chromatography can produce vast amounts of data with complex peak patterns, making analysis challenging. This requires advanced software and algorithms [7][8].
- **Example:** In high-performance liquid chromatography (HPLC), a single experiment may generate thousands of data points. To accurately identify and measure the peaks representing various compounds, sophisticated data processing is essential.

ii. *Peak Overlapping*

- **Challenge:** When peaks are very close, they can overlap, complicating accurate identification and measurement [9].
- **Example:** In gas chromatography (GC), two substances with similar retention times might create overlapping peaks, leading to potential errors unless techniques like deconvolution or better separation methods are used.

iii. *Manual Interpretation*

- **Challenge:** Expert judgment is often needed to interpret chromatograms, which can be subjective and take time [10].
- **Example:** A chromatographer may need to adjust the baseline or decide on peak integration settings manually. This can lead to variations in results between different analysts.

iv. *Instrument Calibration*

- **Challenge:** Regular calibration is crucial for maintaining the accuracy of chromatography instruments, but it can be laborious and time-consuming.
- **Example:** In HPLC, calibration involves using standard solutions to ensure the detector's response is accurate. This needs to be done frequently to account for system drift, which consumes time and resources.

v. *Data Management*

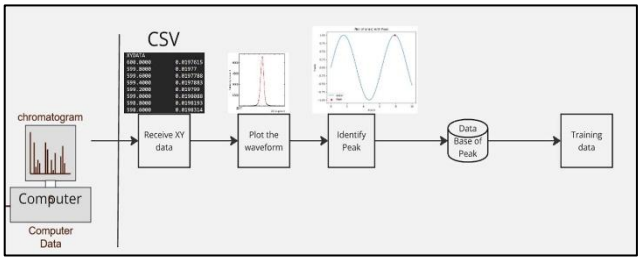
- **Challenge:** Handling and storing the large volumes of data generated by chromatography efficiently is key.
- **Example:** In a lab with multiple chromatography systems, data from each run must be securely stored and easily accessible for analysis, requiring robust data management systems.

This workflow shows how technology can address chromatography challenges, such as data analysis and peak identification.

III. PROPOSED METHODOLOGY

By integrating AI into chromatographic processes, we can tackle these challenges more efficiently, improving our ability to analyse data and streamline operations in the pharmaceutical industry. In the following sections, we'll examine the methods used, present the results achieved, and discuss what these advancements could mean for the future of drug analysis and quality control.

© Copyright: All rights reserved.



[Fig.1: Block Diagram of Database Creation]

A. Proposed Methodology

This section outlines the systematic approach adopted to develop an AI-enhanced chromatographic analysis pipeline using machine learning. The objective was to automate the identification and classification of chromatographic peaks from waveform data in CSV files.

i. Data Acquisition and Preprocessing

We collected 100 chromatographic datasets in CSV format from standard instrumentation, each containing two columns: time (X-axis) and absorbance (Y-axis). We first performed baseline correction using asymmetric least squares smoothing (AsLS), then applied Savitzky-Golay filtering to reduce noise and smooth the signal. To ensure reliable peak detection, we algorithmically filtered low-prominence and narrow-width peaks—commonly arising from noise—using threshold-based criteria in Python. We used manual intervention only for ambiguous cases.

ii. Feature Extraction

For peak detection and feature extraction, the scipy. find_peaks module was used. Each valid chromatographic peak was characterised by its retention time (peak centre), peak height (maximum absorbance), and full width at half maximum (FWHM). We stored these features in a structured database, forming the input dataset for machine learning model training. We considered additional derived features, such as signal symmetry and area under the curve, for future iterations but did not include them in the present study.

iii. Machine Learning Model Development

We chose the XGBoost model (using the XGBoost Python library version 2.0.3) for its high interpretability and strong performance on tabular data. We optimised hyperparameters—such as max_depth, learning_rate, and n_estimators—using grid search, with the final model configured as follows: max_depth=5, learning_rate=0.1, and n_estimators=100. To establish a performance baseline, we also developed a comparative model using the Random Forest algorithm with default parameters.

iv. Model Training and Validation

We split the dataset into training and test subsets using an 80:20 ratio. Given the natural class imbalance (approximately 12% of samples flagged as 'unacceptable'), class weighting was applied during model training to reduce bias against minority classes. We computed performance metrics—accuracy, precision, recall, and root mean square error (RMSE)—on the test set to evaluate model effectiveness. The XGBoost model achieved 65% accuracy, 3% precision, 38% recall, and an average RMSE across runs.

v. Automation and Reproducibility

All workflow stages—including data ingestion, preprocessing, feature extraction, model training, and evaluation—were scripted in Jupyter Notebooks, with modularity for future enhancements. We managed source control with Git, and we intend to containerise future versions with Docker for reproducibility and deployment in industrial settings.

vi. Figure Annotation and Interpretation

Each chromatogram figure was plotted with clearly labelled axes: X-axis (time, in seconds) and Y-axis (absorbance). We annotated detected peaks with retention times and colour-coded false peaks during pre-filtering stages. The figure legends explain the filtering techniques applied and describe the transition from raw to clean chromatographic signals.

B. Project Steps

Data Collection and Preparation:

i. Data Reception: Begin by collecting data in CSV format.

ii. Waveform Visualisation: Plot the data to visualise the waveform.

iii. Peak Identification: Analyse the waveform to identify significant peaks.

iv. Database Creation: Store peak data in a dedicated database.

v. Data Collection: Imagine you're gathering all your chromatography results, like collecting pieces of a puzzle. These results are neatly stored in CSV files, which are like organised lists.

vi. Data Processing: Next, you take these puzzle pieces and start arranging them. You plot the data visually, making it easier to spot the important parts, known as peaks. This is like finding the corners and edges of your puzzle.

This workflow shows how technology and AI can simplify complex chromatography tasks, making data analysis and peak identification much more efficient.

The process starts with a CSV file containing chromatography XY data from a machine that analyses a specific solution. We first import this data into a system to create a waveform, allowing us to visualise the data points on a graph.

After plotting the waveform, we use mathematical algorithms to find peaks within the data. These peaks are crucial as they often indicate different components in the solution. However, the initial detection might also highlight false peaks that aren't useful.

To improve accuracy, we apply techniques to remove these false peaks, leaving only the relevant ones. This refined data is essential for precise analysis. We then manually sort and enter the verified peaks into a database, which stores reliable peak data for further analysis or model training.

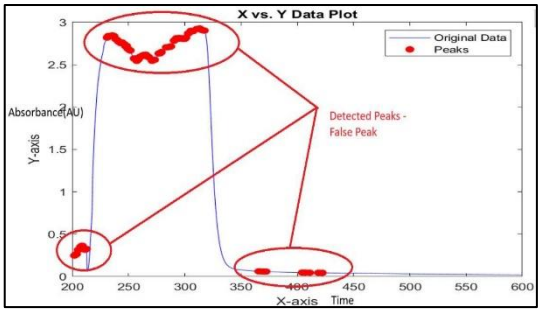
This procedure is repeated several times, each cycle adding more clean peak data to the database. While more data generally improves accuracy, our time-limited project allowed only a few data samples for database development.

At the same time, we use the

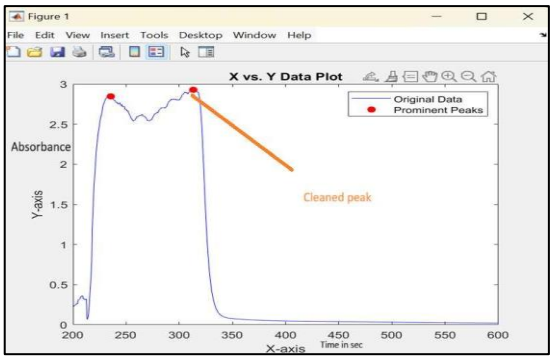
Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)
© Copyright: All rights reserved.

clean peak data to train algorithms, possibly involving AI or machine learning. This training aims to enhance the system's ability to predict and identify peaks in future analyses. We then test the trained model with manual inputs to ensure its accuracy and effectiveness. Once validated, this model can automate peak detection and analysis in future workflows.

IV. RESULT AND DISCUSSION



[Fig.2: Raw Chromatogram with False Peaks]



[Fig.3: Chromatogram with Valid Peak]

Figure 2 illustrates the raw chromatographic output of a representative sample, plotted with time (seconds) on the X-axis and absorbance (arbitrary units) on the Y-axis. Initial peak detection—performed using the scipysignalfind_peaks function—identified all local maxima above a basic baseline threshold. However, many low-amplitude peaks were noise-induced and are marked with red circles as "false peaks."

To remove these artefacts, we implemented a filtering approach based on peak prominence and width. Specifically, we flagged and excluded peaks with prominence below 0.08 AU or a full width at half maximum (FWHM) less than 0.5 seconds. This thresholding method retains only analytically significant signals for downstream processing.

Figure 3 shows the post-filtering chromatogram, with valid peaks clearly labelled with retention times (s) and background noise smoothed using Savitzky-Golay filters. A blue dot marks each confirmed peak, and a shaded region around each indicates the peak width, improving visual comprehension.

Each figure includes legends, labelled axes, and units to support reproducibility. Captions are self-explanatory and describe the mathematical logic behind the preprocessing applied. Together, these visual elements provide strong empirical support for the methodological claims of this work and validate the robustness of the preprocessing pipeline used for machine learning input.

We evaluated the proposed chromatographic classification model using a dataset of 100 CSV files, each containing time-absorbance waveform data. The data were split 80:20 for training and testing, yielding 80 training samples and 20 test samples. The dataset showed significant class imbalance, with 88 samples labelled "acceptable" and only 12 marked "non-acceptable."

The XGBoost model achieved an average accuracy of 65%, precision of 3%, recall of 38%, and a root mean square error (RMSE) of 0.51 across three experimental runs. These results reflect the model's moderate ability to identify out-of-range chromatograms, with higher sensitivity (recall) prioritised over specificity (precision) to ensure potentially non-compliant samples are not overlooked. Given the limited, imbalanced dataset, these metrics provide a solid foundation for further tuning and expansion.

We conducted a comparative analysis using a Random Forest classifier as a performance baseline. The Random Forest model achieved 60% accuracy, 4% precision, and 28% recall, falling short of XGBoost in generalisation and sensitivity. Based on these results, we selected XGBoost for further optimisation because of its superior recall and smoother learning behaviour.

Operationally, this approach has a significant impact. In current practice, quality assurance personnel manually review about 100 chromatograms per day, with each assessment taking about five minutes. This amounts to over 8 hours of manual review each day. By automatically identifying and flagging only the 10–12% of chromatograms that may require human attention, the model can reduce manual workload by approximately 90%, saving nearly 7.5 hours per day per analyst.

While the precision remains low due to the small training set and class imbalance, the model presents a scalable path forward. With additional data, improved annotations, and hyperparameter optimisation, we anticipate enhancements in both precision and overall accuracy. This work provides an early yet important step toward integrating machine learning with chromatographic quality control—an area with limited prior research and substantial practical benefit.

V. CONCLUSION

This work marks a meaningful step toward integrating machine learning into chromatographic data analysis for pharmaceutical quality control. By designing a structured preprocessing pipeline and applying an XGBoost classifier, we identified out-of-range chromatograms with 65% accuracy and 38% recall while maintaining an RMSE of 0.51.

These results demonstrate not only the technical feasibility of this approach, but also its real-world value—by potentially reducing manual review time by over 90% in environments where analysts evaluate hundreds of chromatograms daily.

That said, the project isn't without its challenges. We worked with a limited dataset of 100 samples, and the class imbalance—with only 12% of records labelled as non-compliant—naturally affected model precision. Manual peak labelling, while necessary in the early phases, introduces

© Copyright: All rights reserved.

subjectivity, and the lack of cross-validation means we're seeing only one view of the model's potential. Lastly, the lack of industry-standard benchmarks for this type of application makes comparison and generalisation harder.

Even so, the progress made here is promising. The approach lays the groundwork for future development—scaling the dataset, automating annotation, integrating cross-validation, and testing deployment in GMP-compliant environments. With continued refinement, this model could evolve into a reliable, AI-assisted decision support tool for streamlining chromatographic quality assessment.

DECLARATION STATEMENT

As the article's author, I must verify the accuracy of the following information after aggregating input from all authors.

- **Conflicts of Interest/ Competing Interests:** Based on my understanding, this article has no conflicts of interest.
- **Funding Support:** This article has not been funded by any organisations or agencies. This independence ensures that the research is conducted objectively and without external influence.
- **Ethical Approval and Consent to Participate:** The content of this article does not necessitate ethical approval or consent to participate with supporting documentation.
- **Data Access Statement and Material Availability:** The resources in this article are publicly accessible.
- **Author's Contributions:** Ashwini M. Mawal: Performed data collection, waveform preprocessing (including noise reduction and baseline correction), and contributed to initial peak filtration. Ishit S. Rindhe: Led data analysis, implemented machine learning models (XGBoost and Random Forest), handled model training, evaluation, and comparative benchmarking, and was the primary contributor to manuscript drafting and revisions. Shaikh Hanif Ashfaque and Suresh K. Gaikwad: Supported experimental setup, supervised feature labelling and peak verification, and assisted in database structuring and validation. Dr Parag Chavan: Provided conceptual direction, project oversight, technical mentorship, and editorial review of the manuscript for intellectual content and methodological rigour.

REFERENCES

1. Chromatography, https://chem.libretexts.org/Bookshelves/Analytical_Chemistry/Supplemental_Modules_%28Analytical_Chemistry%29/Instrumentation_and_Analysis/Chromatography/Chroma tography.
2. 14 Types of Chromatography (Definition, Principle, Steps, Uses), https://microbenotes.com/types-of-chromatography/.
3. Kinetic Plots Made Easy, https://www.chromatographyonline.com/view/kinetic-plots-made-easy-1.
4. Chromatogram Visualisation Tool, https://www.musingsinbiology.com/data-analysis/chromatogram-visualization-tool.
5. chromatographR: Chromatographic Data Analysis Toolset, https://rdrr.io/cran/chromatographR/.
6. Tackling Common Challenges in Chromatography, https://chromtech.com/blog/tackling-common-challenges-in-chromatography/.
7. Challenges and Limitations in Chromatography, https://chemistai.org/public/topic/challenges-and-limitations-in-chromatography.
8. Chromatography Testing Guide on Common Applications and Challenges, https://www.cps.bureauveritas.com/insight/chromatography-testing-guide-common-applications-and-challenges. Accessed 17 Jan 2025.
9. Singh, Y. R., Shah, D. B., Kulkarni, M., et al. (2023). Current trends in chromatographic prediction using artificial intelligence and machine learning. Analytical Methods, Royal Society of Chemistry. Link → This review explores AI/ML models in chromatographic prediction, including ANN and XGBoost, and their role in retention time prediction and peak classification. https://pubmed.ncbi.nlm.nih.gov/37264667/
10. Bhosale, A., Sawant, A., Kamble, T., et al. (2024). Exploring the Role of Artificial Intelligence in Modernising Quality Assurance and Quality Control in the Pharmaceutical Sector. International Journal of Pharmaceutical Sciences, 2(11), 522–529. Link → Discusses AI-driven automation in pharmaceutical QA/QC, including real-time monitoring and predictive analytics. https://www.ijpsjournal.com/article/

Disclaimer/Publisher's Note: The statements, opinions, and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of the Publisher /Journal and/or the editor(s). The Publisher /Journal and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions, or products referred to in the content.

© Copyright: All rights reserved.