ElectroML: An Open-Source Software for Machine Learning-Based Analyte Concentration Prediction in Electrochemical Sensing

Canan Hazal Akarsu[a,*], Tarık Küçükdeniz[a*], Elif Tüzün[b,c], Selcan Karakuş[b,c]

[a]Istanbul University-Cerrahpaşa, Faculty of Engineering, Department of Industrial Engineering, 34320, Istanbul, Türkiye
[b]Istanbul University-Cerrahpaşa, Faculty of Engineering, Department of Chemistry, 34320, Istanbul, Türkiye
[c]Health Biotechnology Joint Research and Application Center of Excellence, Esenler, 34220 Istanbul, Türkiye

Abstract

ElectroML is an open-source machine learning (ML) software designed to transform electrochemical data analysis and prediction. The software automates the processing of cyclic voltammetry (CV) and differential pulse voltammetry (DPV) data, integrating six state-of-the-art ML models including Artificial Neural Networks, Support Vector Machines, and gradient boosting methods. Beyond training models on known datasets, ElectroML enables rapid prediction of analyte concentrations from new sensor measurements, making it valuable for environmental monitoring, medical diagnostics, and industrial quality control applications. Its interactive Streamlit interface allows researchers without extensive programming experience to train models and predict concentrations. Its modular, domain-independent design ensures adaptability across scientific domains.

Keywords

Machine Learning, Electrochemical Analysis, Voltammetry, Concentration Prediction, Artificial Neural Networks, XGBoost

Metadata

Nr	Code metadata description	*Metadata*
C1	Current code version	*V1.0*
C2	Permanent link to code/repository used for this code version	*https://github.com/tkucukdeniz2/ElectroML*
C3	Permanent link to reproducible capsule	*None*
C4	Legal code license	*GNU General Public License (GPL)*
C5	Code versioning system used	*Git*
C6	Software code languages, tools and services used	*python*
C7	Compilation requirements, operating environments and dependencies	*streamlit>=1.31.0, pandas>=2.2.0, numpy>=1.24.0, scipy>=1.12.0, scikit-learn>=1.4.0, xgboost>=2.0.0, lightgbm>=4.2.0, tensorflow>=2.15.0, matplotlib>=3.8.0, seaborn>=0.13.0, plotly>=5.18.0, openpyxl>=3.1.0, xlsxwriter>=3.1.0*

*Corresponding Author: Canan Hazal Akarsu
E-mail: hazalakarsu@iuc.edu.tr

		protobuf>=4.25.0
C8	If available, link to developer documentation/manual	*https://github.com/tkucukdeniz2/ElectroML/blob/main/README.md*
C9	Support email for questions	hazalakarsu@iuc.edu.tr, tkdeniz@iuc.edu.tr

1. Motivation and significance

The detection and quantification of analytes are vital for scientific research, ecosystem preservation, environmental sciences and public health. Numerous biochemical and environmental processes depend on the presence of analytes such as heavy metals, pesticides, drugs, nutrients (e.g. phosphates and nitrates) and biomarkers, and the presence of these substances outside certain concentrations can pose serious risks. Accurate monitoring is essential in environmental, agricultural, and clinical applications. While traditional detection methods are effective, they are often complex and resource-intensive. Electrochemical methods, such as differential pulse voltammetry (DPV) and cyclic voltammetry (CV), offer effective, real-time measurements with high sensitivity, specificity, and selectivity. They are also cheap and flexible for portable, on-site monitoring, and they offer fast detection, low detection limits, good reproducibility, and the capacity to differentiate analytes in complex matrices.

Electrochemical sensors are effective tools for detecting environmental pollutants at low concentrations. In 2022, Shanbhag et al. developed a hafnium-doped tungsten oxide ($Hf.WO_3$)-based carbon paste electrode for the detection of perfluorooctanoic acid (PFOA), achieving high sensitivity (detection limit: 1.83×10^{-8} M) and selectivity under neutral pH and optimal temperature [1]. The electrode performed well in environmental samples, unaffected by metal ions. Similarly, Solís et al. used a gold nanoparticle (Au NPs)-coated glassy carbon electrode modified with 1H,1H,2H,2H-Perfluorodecanethiol (PFDT) to detect PFOA via SW-AdCSV, achieving limits of 24 ppt (detection) and 80 ppt (quantification) [2]. In another study, Dash et al. developed ultrasensitive electrochemical sensors using copper and copper@silver nanorods for simultaneous heavy metal detection [3]. The sensors exhibited low detection limits, high sensitivity, and good reproducibility, with over 90% recovery in water samples. They offered low-cost, repeatable alternatives to commercial screen-printed electrodes. These electrochemical studies highlighted the potential of low-cost, portable sensors for environmental monitoring and water quality analysis.

Despite their advantages, interpreting voltammetric data is challenging due to the influence of factors such as pH, ionic strength, and electrode surface properties. These complexities hinder the accurate extraction of signal features corresponding to analyte concentrations. Traditional statistical methods often struggle with the nonlinear relationships in such data, limiting their predictive accuracy. ML offers a transformative approach to improving the predictive accuracy of electrochemical sensing. ML models effectively identify patterns in high-dimensional, nonlinear datasets, making them ideal for analyzing complex signals. Integrating ML enhances the precision of analyte concentration predictions, even under noisy or variable conditions. Softwares such as TSFEL [4], which streamline the feature extraction process for time-series data, and NiaAML [5], an AutoML software for optimizing classification and regression pipelines, exemplify the potential of ML in handling complex datasets efficiently.

The development of a generalized ML software for electrochemical data analysis addresses several critical needs:

1. Automation and Efficiency: Manual feature extraction and model development are time-consuming and prone to human error. A standardized ML software automates these processes, ensuring consistent and reproducible results across various datasets and experimental conditions. This automation accelerates research workflows and reduces errors through automated feature extraction.

2. Predictive Capabilities: The software enables concentration prediction from new voltametric measurements using ML models trained on known sensor data. Researchers can train models using datasets with known concentrations, then use these models to predict concentrations from new sensor readings. This capability is particularly valuable for rapid analysis in environmental monitoring and quality control applications.
3. Improved Predictive Power: By leveraging advanced ML algorithms such as artificial neural networks (ANN), support vector machines (SVM), and gradient boosting methods, the software outperforms traditional regression methods in predictive accuracy. This improvement is especially crucial for detecting trace analyte concentrations, where even small inaccuracies can have significant implications for environmental, agricultural, or health-related assessments.
4. Scalability and Flexibility: The integration of Leave-One-Out Cross-Validation (LOOCV) and multiple modeling options ensures robust performance across different dataset sizes. The software's web interface allows users to easily train models and make predictions, making it accessible for both research and practical applications.
This ML software offers a reliable solution for analyte detection, transforming electrochemical sensing through automation and accurate prediction capabilities. It enables rapid concentration determination from new sensor measurements, accelerating progress in environmental monitoring, water quality analysis, and medical diagnostics. The combination of automated preprocessing, multiple modeling options, and prediction capabilities makes it a valuable tool for both research and practical applications.
ML has significantly enhanced the analysis of electrochemical data, enabling precise analyte detection across various applications. For instance, Kayali et al. demonstrated Random Forest's effectiveness in analyzing potassium ferrocyanide concentrations using DPV and SQWV, achieving 100% accuracy and low detection limits, highlighting ML's potential for environmental monitoring [6]. DPV techniques, known for their high sensitivity in detecting low-concentration analytes, were integrated with ML by Zhang et al. to develop a glucose detection model [7]. XGBoost outperformed traditional methods, reducing error by 25.3% and improving accuracy over a broader data range. This approach highlights ML's potential for mobile and real-time sensing systems in biomedical and environmental monitoring. ML algorithms such as RF, SVM, and Gradient Boosting, excel at enhancing electrochemical signal analysis. Naghian et al. demonstrated RF's superior accuracy, even with limited data, by employing noise-reduction preprocessing and Optuna for hyperparameter tuning [8]. This versatility supports applications in environmental monitoring, diagnostics, and food safety. In biomedical diagnostics, Kammarchedu et al. developed an ML model for eMoSx-LIG-based sensors, enabling the simultaneous detection of biomarkers like tyrosine and uric acid with higher accuracy and lower detection limits [9].
This multimodal approach promises significant advancements in environmental monitoring and food safety by addressing complex electrochemical signals. Environmental applications also benefit from the integration of ML. Meskher et al. combined reduced graphene oxide (rGO) and metal-organic frameworks (MOFs) to develop a sensor for pentachlorophenol detection [10]. Using an ANN model, they achieved an LOD of 75.63 nM, highlighting ML's ability to improve sensor accuracy and reliability. Wearable ML-integrated sensors are advancing biomedical diagnostics. Bao et al. developed a graphene oxide(GO) -carbon black sensor for tyrosine detection, utilizing ANN and SVM algorithms to achieve an R^2 value of 0.9828 [11]. These portable, low-cost sensors offer sensitive, noise-reduced analysis for biological and environmental samples. Finally, ML addresses critical challenges such as temperature drift in electrochemical sensors. Bhardwaj et al. integrated a Simulation Program with Integrated Circuit Emphasis (SPICE) macromodel with Random Forest (RF) to compensate for temperature variations in pH measurements, demonstrating ML's role in sensor design for environmental and biomedical applications [12].

2. Software description

This software is designed to facilitate the prediction of analyte concentrations using CV and DPV data. By integrating advanced ML algorithms and robust feature extraction techniques, the software automates the analysis of complex electrochemical datasets. The modular architecture ensures adaptability to a variety of experimental conditions, enabling researchers to extract meaningful insights efficiently. The software's open-source nature fosters reproducibility and scalability, making it an invaluable tool for scientific discovery and practical applications.

2.1. Software architecture*:*

The ElectroML software is designed with a modular architecture to ensure flexibility, scalability, and ease of use across diverse datasets and research objectives. The software is structured into four modules:

Data Preprocessing Module: This module is responsible for feature extraction, normalization, and dataset preparation. It supports multiple file formats such as Excel and CSV, ensuring compatibility with CV and DPV datasets. Advanced statistical and geometrical features are extracted from sensor data, including peak counts, skewness, kurtosis, and positive area under the curve. The module also calculates feature importance using a Random Forest model and analyzes correlations with target variables, providing insights into dataset characteristics. All outputs are saved in CSV format and visualized for clear interpretation.

Model Training Module: This module implements a range of ML algorithms, including Artificial Neural Networks (ANN), Support Vector Machines (SVM), Random Forest (RF), XGBoost, LightGBM, and Linear Regression (LR). It employs LOOCV for robust model evaluation. Performance metrics such as Mean Squared Error (MSE), Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R^2 Score are computed to assess model effectiveness. Hyperparameter tuning is seamlessly integrated using the Optuna library, enabling optimized model performance tailored to specific datasets. All model predictions and metrics are stored in CSV files for reproducibility.

Visualization Module: This module generates detailed plots to aid in the interpretation and communication of results. Key visualizations include Actual vs. Predicted scatter plots and Residual Plots. These visualizations provide a clear understanding of model performance and highlight potential areas for improvement. All plots are saved as publication-quality images.

Prediction Module: This module enables concentration prediction from new sensor measurements using trained ML models. It supports the same data formats and preprocessing steps as the training pipeline, ensuring consistency between model training and prediction. Users can upload new voltametric measurements, and the module automatically extracts features using the same preprocessing algorithms. The selected trained model, along with its corresponding scaling parameters, is then used to predict analyte concentrations. The module provides both numerical predictions and visual representations of the results, with options to export predictions for further analysis. This streamlined prediction workflow makes the software practical for routine analysis and real-time monitoring applications.

2.2. Software functionalities*:*

The software offers a set of functionalities to address challenges in analyzing electrochemical data:

- Data Preprocessing: Efficiently handles feature extraction, normalization, and dataset preparation.
- Model Training and Optimization: Implements advanced ML models, including ANN, SVM, RF, XGBoost, LightGBM, and LR, with hyperparameter tuning via Optuna.
- Prediction: Enables accurate prediction of analyte concentrations from new sensor data using trained ML models.

- Evaluation Metrics: Calculates critical metrics such as MSE, MAE, RMSE, and R^2 Score for robust performance evaluation.
- Visualization: Generates plots like actual versus predicted values and residuals to interpret and communicate results effectively.
- Results Export: Saves predictions, metrics, and visualizations in user-friendly formats such as CSV and PNG.

These functionalities ensure that the software addresses the complexities of voltammetric signal analysis while providing a streamlined workflow for researchers.

ElectroML is also accessible as an interactive Streamlit application to streamline electrochemical data analysis. This web-based platform allows users to process data, train ML models, visualize results and make predictions all within a user-friendly interface. The application is hosted at the following link: https://electroml.streamlit.app/.

3. Illustrative examples

The following example demonstrates the full potential of the ElectroML Streamlit application. By leveraging its intuitive interface, users can perform electrochemical data analysis with ease, focusing on DPV data for analyte concentration prediction. This example walks through the four key modules of ElectroML: Data Preprocessing, Model Training, Results Visualization and Prediction.

The first step in the analysis involves preparing the sensor data for ML. The user uploads the sensor_data.xlsx file via the Data Preprocessing module in the web interface. The dataset used in this example is adapted from Gürsu et al. [13] The platform automatically displays a preview of the dataset, including basic statistical summaries such as the total number of samples and features. The user initiates the feature extraction process by clicking the "Feature Extraction Process" button. The platform computes a variety of features, such as mean, median, kurtosis, and number of peaks, and displays the extracted features in a table format. The extracted features are analyzed for their importance in predicting the target variable (analyte concentration). A bar chart ranks features based on their relative importance, helping users identify the most significant predictors. Extracted features are saved as processed_features.csv. Feature importance plot is generated for better interpretation.

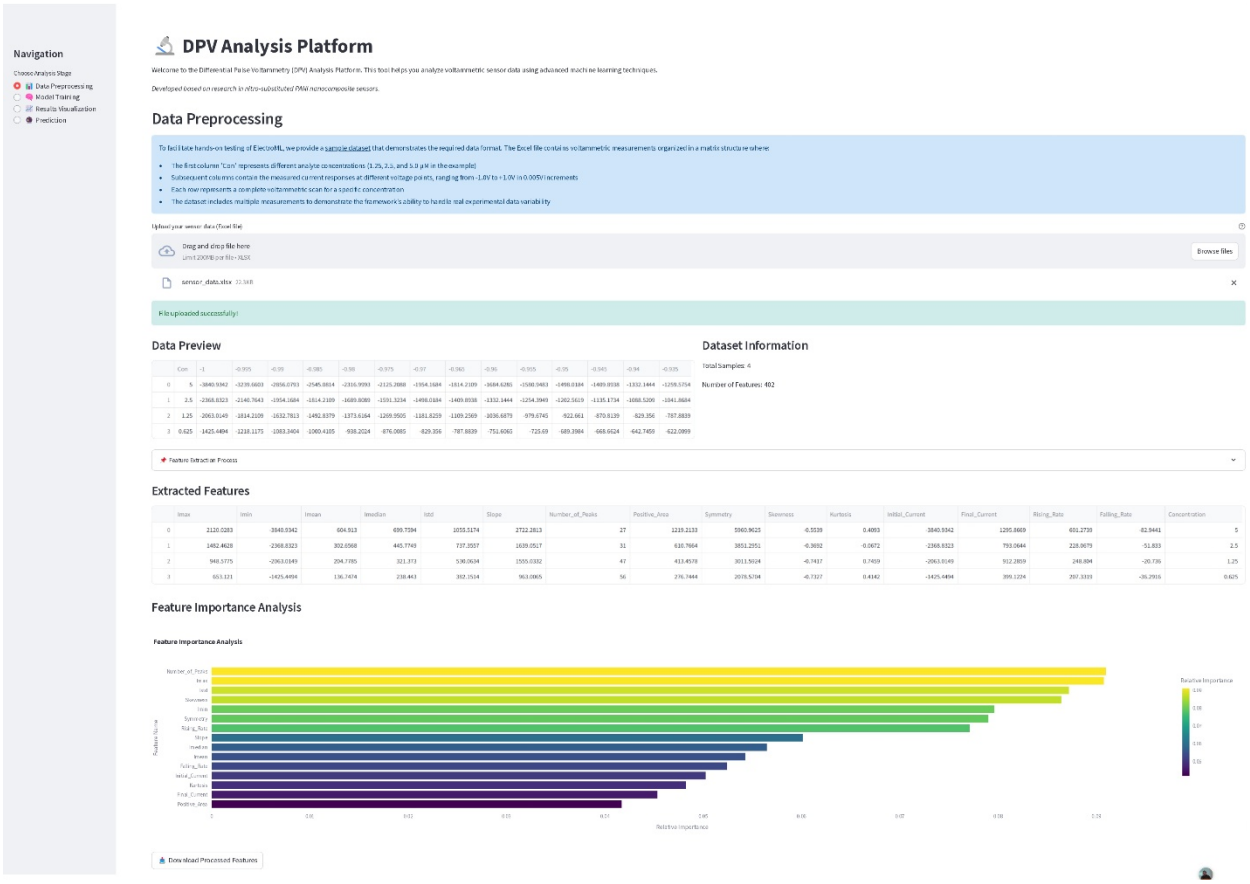

Figure 1: Data Preprocessing Module Interface

After preprocessing the data, the next step is to train ML models. The processed_features.csv file is uploaded in the Model Training module, where the platform confirms a successful upload and displays a summary of the features. Users can select ML models from a list, including Linear Regression, SVM, Random Forest, XGBoost, LightGBM, and ANN. Multiple models can be selected simultaneously for training. By clicking the "Train Selected Models" button, the platform trains the chosen models using the processed features. Progress indicators show the completion status of each model, and the performance metrics are summarized in a table. Performance metrics are displayed for each model. Model-specific results are saved as CSV files.

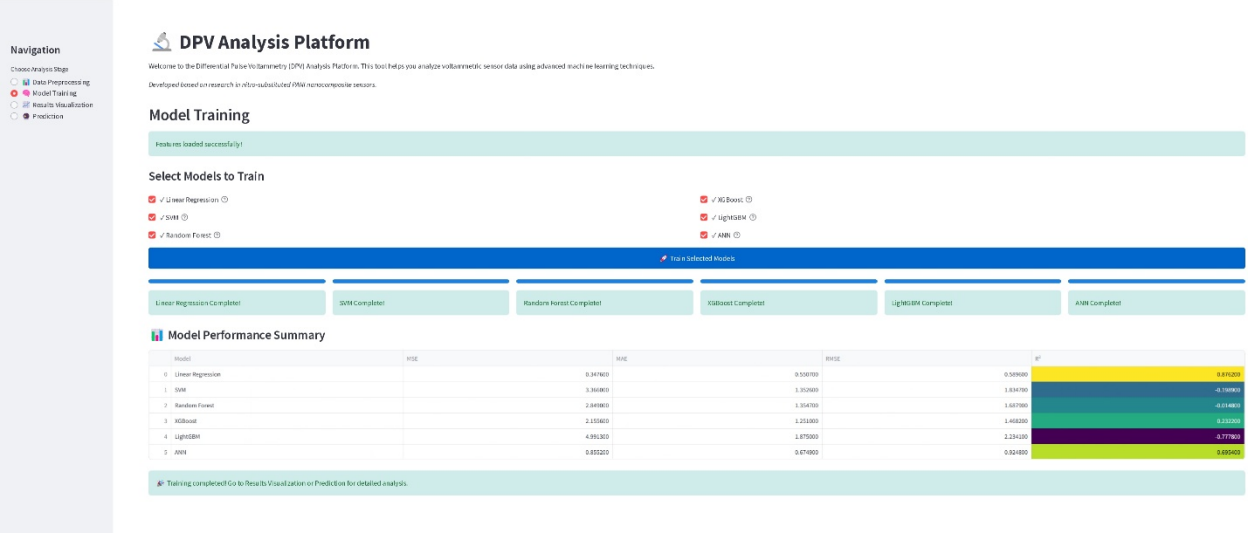

Figure 2: Model Training Module Interface

The Results Visualization module provides an overview of model performance through bar charts. Users can compare models based on metrics like R^2 Score to identify the best-performing algorithm. Users can select a specific model and visualization type for deeper analysis. For example, the platform can display Actual vs Predicted Plot, Residual Plot or Error Distribution Plot. Final results and visualizations can be exported as CSV and PNG files, enabling further offline analysis.

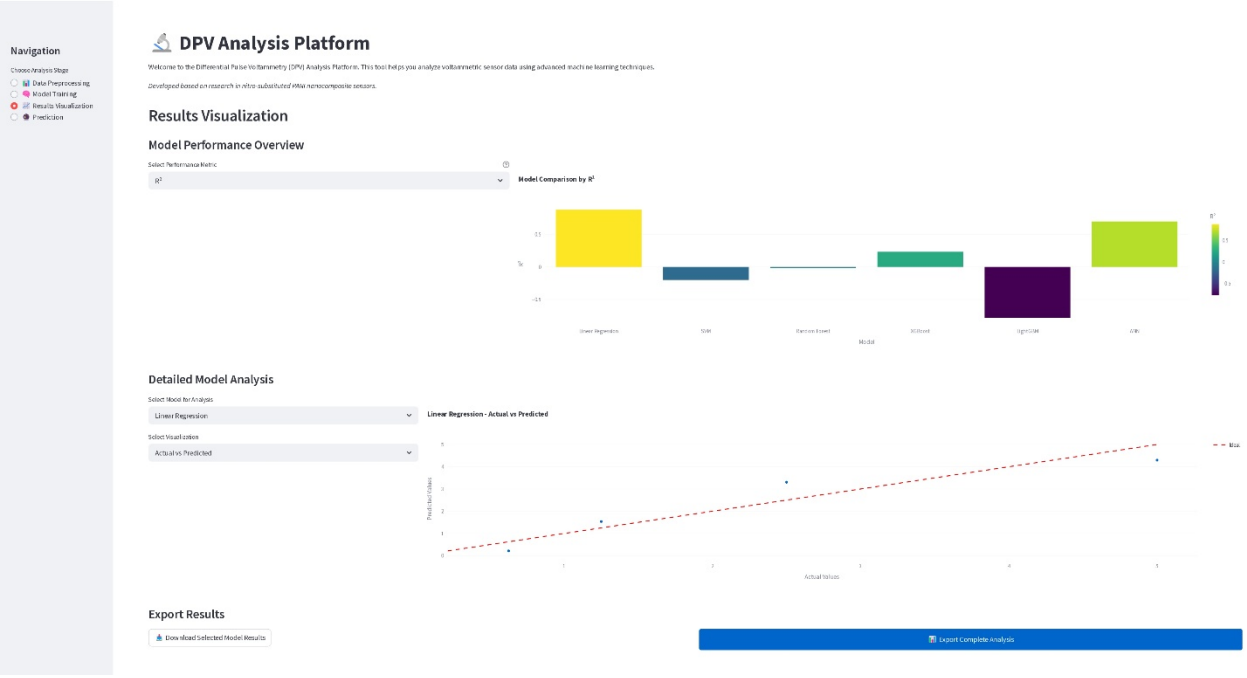

Figure 3: Results Visualization Module Interface

The Prediction module is the final step in the ElectroML workflow, enabling users to predict analyte concentrations from new DPV sensor data. This module builds on the models trained in the earlier steps to provide accurate and reliable concentration predictions. Users begin by uploading a new DPV dataset (e.g., data.xlsx) that follows the same format as the training data. Users can see the extracted features by clicking the "Feature Extraction Process" button. The platform allows users

to select one of the previously trained models. The chosen model will be used to predict analyte concentrations based on the extracted features. By clicking the "Predict Concentration" button, the platform computes the predicted concentrations for each sample in the dataset. The predictions are displayed in a table format alongside a scatter plot that visualizes predicted concentrations by sample. The module also provides a summary of model performance metrics. Users can download the prediction results and performance metrics as Excel files for further analysis or documentation by clicking the "Download Predictions" button.

This example illustrates the comprehensive capabilities of ElectroML, showcasing its ability to preprocess DPV data, train multiple ML models, generate detailed visual insights, and accurately predict analyte concentrations. The addition of the Prediction module highlights the practical utility of ElectroML in real-world scenarios, providing an end-to-end solution for electrochemical data analysis. With its automated workflow, the framework ensures a user-friendly experience while delivering reliable and reproducible results.

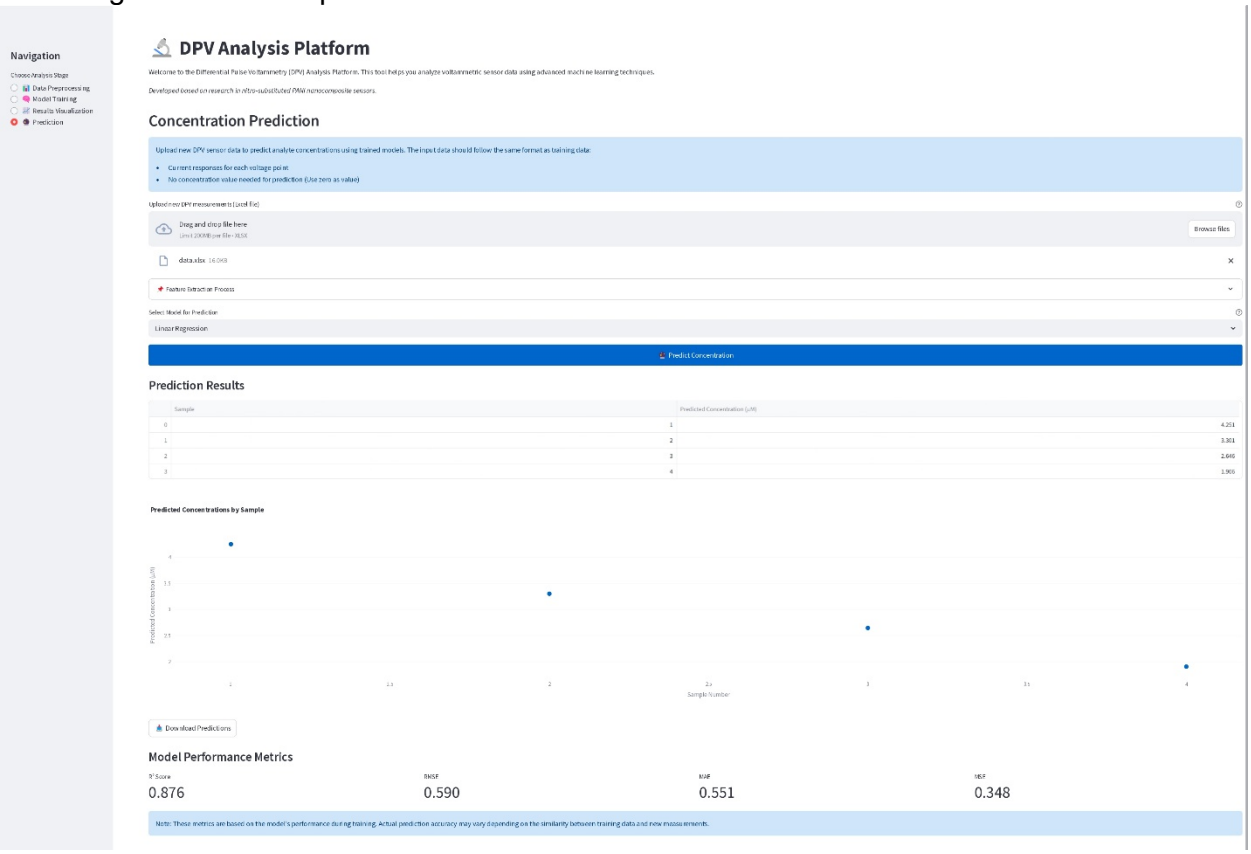

Figure 4: Prediction Module Interface

4. Impact

ElectroML can be used in the investigation of voltammetric feature patterns across different electrode materials, optimization of sensor design parameters through automated feature analysis, and exploration of structure-property relationships in electrochemical sensors. The software's ability to handle diverse datasets enables researchers to study cross-material comparisons and develop novel electrode modifications.

The software significantly improves existing electrochemical research practices by automating time-consuming analysis steps. Traditional voltammetric data analysis requires extensive domain expertise and manual processing, often taking hours for complex datasets. ElectroML reduces this to minutes while minimizing human error. The integration of advanced ML algorithms like

XGBoost and ANN improves prediction accuracy compared to conventional statistical methods, as demonstrated in [13] where the software was used to analyze the impact of nitro substituents on dopamine sensing using PANI nanocomposite sensors.

ElectroML has transformed the daily workflow of electrochemical researchers by automating feature extraction, enabling rapid model training, and providing concentration predictions from new measurements. Through its interactive interface, researchers without extensive programming experience can now implement advanced ML techniques. The standardized analysis pipelines ensure reproducibility across different laboratories, while the web-based deployment ensures accessibility across platforms.

The software's open-source nature and web-based interface have facilitated its adoption in academic research, as evidenced by its application in electrochemical sensing studies [13]. Its capabilities make it particularly suitable for environmental monitoring systems, quality control in industrial processes, and clinical diagnostics applications requiring standardized analysis procedures.

Through its combination of automated workflows, prediction capabilities, and user-friendly interface, ElectroML establishes a foundation for advancing both research and practical applications in electrochemical sensing. Its modular architecture ensures adaptability to emerging research needs, while the open-source framework encourages collaborative improvement and extension to new application domains.

5. Conclusions

This study introduces an open-source ML software designed for the prediction of analyte concentrations using CV and DPV data. The software addresses critical challenges in electrochemical data analysis by automating feature extraction, model training, and performance evaluation. Its modular architecture and integration of advanced ML algorithms enable researchers to process complex datasets efficiently and achieve high predictive accuracy.

The proposed software not only improves existing workflows but also opens new avenues for research. By streamlining data analysis and providing robust tools for optimization, it facilitates the development of more sensitive sensors and the exploration of cross-disciplinary applications, such as materials science and environmental monitoring. The Streamlit application further enhances usability, allowing researchers and educators to engage with the software through an intuitive, web-based interface, thereby promoting accessibility and ease of use. Additionally, the software's open-source nature ensures reproducibility and fosters collaboration among researchers, paving the way for future enhancements and broader adoption. The combination of open access and a user-friendly platform empowers the community to contribute to its development, enhancing its functionality and expanding its potential applications.

As this marks the first release of the software, future developments will focus on incorporating additional ML techniques, expanding its functionalities, and improving usability based on community feedback. By bridging the gap between traditional electrochemical methods and modern ML approaches, and by leveraging the interactive capabilities of the Streamlit application, this software is poised to become an essential tool for advancing scientific discovery and practical applications in electrochemical sensing.

References

[1] Shanbhag MM, Shetti NP, Kalanur SS, Pollet BG, Nadagouda MN, Aminabhavi TM. Hafnium doped tungsten oxide intercalated carbon matrix for electrochemical detection of perfluorooctanoic acid. Chem Eng J. 2022;434:134700.

[2] Solís JJC, Yin S, Galicia M, Ersan MS, Westerhoff P, Villagrán D. "Forever chemicals" detection: A selective nano-enabled electrochemical sensing approach for

perfluorooctanoic acid (PFOA). Chem Eng J. 2024;491:151821. https://doi.org/10.1016/j.cej.2022.134700.
[3] Dash SR, Bag SS, Golder AK, Ivaturi A. Ultrasensitive electrochemical sensors based on Cu and Cu@ Ag nanorods for simultaneous heavy metal detection. Mater Chem Phys. 2024;318:129255. https://doi.org/10.1016/j.matchemphys.2024.129255.
[4] Barandas M, Folgado D, Fernandes L, Santos S, Abreu M, Bota P, et al. TSFEL: Time series feature extraction library. SoftwareX. 2020;11:100456. https://doi.org/10.1016/j.softx.2020.100456.
[5] Fister Jr I, Farthofer LA, Pečnik L, Fister I, Holzinger A. NiaAML: AutoML for classification and regression pipelines. SoftwareX. 2025;29:101974. https://doi.org/10.1016/j.softx.2024.101974.
[6] Kayali D, Shama NA, Asir S, Dimililer K. Machine learning-based models for the qualitative classification of potassium ferrocyanide using electrochemical methods. J Supercomput. 2023;79(11):12472-91. https://doi.org/10.1007/s11227-023-05137-y.
[7] Zhang B, Zhang Y, Shen J, Zhou Z, Zhu G. Research on differential pulse voltammetry detection method for low concentration glucose based on machine learning model. Int J Electrochem Sci. 2024;19(2):100479. https://doi.org/10.1016/j.ijoes.2024.100479.
[8] Naghian E, Marzi Khosrowshahi E, Sohouli E, Pazoki-Toroudi HR, Sobhani-Nasab A, Rahimi-Nasrabadi M, Ahmadi F. Electrochemical oxidation and determination of antiviral drug acyclovir by modified carbon paste electrode with magnetic CdO nanoparticles. Front Chem. 2020;8:689. https://doi.org/10.3389/fchem.2020.00689.
[9] Kammarchedu V, Butler D, Ebrahimi A. A machine learning-based multimodal electrochemical analytical device based on eMoSx-LIG for multiplexed detection of tyrosine and uric acid in sweat and saliva. Anal Chim Acta. 2022;1232:340447. https://doi.org/10.1016/j.aca.2022.340447.
[10] Meskher H, Achi F, Ha S, Berregui B, Babanini F, Belkhalfa H. Sensitive rGO/MOF based electrochemical sensor for penta-chlorophenol detection: a novel artificial neural network (ANN) application. Sens Diagn. 2022;1(5):1032-43. https://doi.org/10.1039/d2sd00100d.
[11] Bao Q, Li G, Cheng W, Yang Z, Qu Z, Wei J, Lin L. Machine learning-assisted flexible wearable device for tyrosine detection. RSC Adv. 2023;13(34):23788-95. https://doi.org/10.1039/d3ra02900j.
[12] Bhardwaj R, Sinha S, Sahu N, Majumder S, Narang P, Mukhiya R. Modeling and simulation of temperature drift for ISFET-based pH sensor and its compensation through machine learning techniques. Int J Circuit Theory Appl. 2019;47(6):954-70. https://doi.org/10.1002/cta.2618.
[13] Gürsu G, Yıldız DE, Taşaltın N, Baytemir G, Karakuş S, Karaca B, et al. Impact of Nitro Substituents on Dopamine Sensing and Nanostructure Morphology: A Machine Learning Approach for PANI: 2-and 3-Nitro-1H-Pyrrole Nanocomposite Sensors. J Electrochem Soc. 2024;171(12):127512. https://doi.org/10.1149/1945-7111/ad9ccb.