DATA ASSIMILATION VS FEEDING AI WITH NEW OBSERVATIONS

Nisryne El Massafi, Chayma El Bacha, Atika Wamra

Computer Systems Engineering Laboratory (LISI), Faculty of Sciences Semlalia, Cadi Ayyad University, Marrakesh, Morocco.

Abstract

Data assimilation techniques are critical for improving the performance of artificial intelligence (AI) models by integrating multi-source data. This study delves into the application of an advanced Kalman filter algorithm for data assimilation with the aim of enhancing the predictive accuracy of AI models. We have developed an AI model and utilized observational data from various sources to refine predictions through the Kalman filter. The research encompasses three main phases: the initial training of the AI model, data assimilation via the Kalman filter, and performance evaluation. Our results indicate an enhancement in model performance, with a substantial reduction in prediction errors, such as Mean Absolute Error (MAE) and an increase in the coefficient of determination (R^2). Notably, the iterative refinement of the AI model's outputs by the Kalman filter ensures the assimilation of new data and the adaptability of the model to changing conditions. The study concludes that data assimilation is an effective approach to enhancing the performance of AI models, especially in systems where real-time data integration is crucial. This work lays the groundwork for further research into the hybridization of AI models with data assimilation techniques and their applications across various fields.

Keywords: Data assimilation, Model performance, Kalman filter, Artificial intelligence (AI) models

1 Literature Review

1.1 Data Assimilation: Overview

1.1.1 Definition

Data assimilation is a potent computational technique widely used in scientific modeling, particularly in atmospheric, oceanic, and land surface studies. It seamlessly integrates observational data with numerical models to enhance prediction accuracy. This approach optimally combines real-world observations and underlying dynamic principles, yielding an updated and more informed representation of the system under study. Still, for this, the freshness and relevance of data with time are significant. Therefore, a model can only offer better results with new and timely observations. Thus, data assimilation compares the new data with the old data and removes relative errors and mistakes, eventually leading to better prediction or forecasting.

1.1.2 Approaches to data assimilation

There are two basic approaches to data assimilation: sequential assimilation, that only considers observation made in the past until the time of analysis, which is the case of real-time assimilation systems, and non-sequential, or retrospective assimilation, where observation from the future can be used, for instance in a reanalysis exercise. Another distinction can made between methods that are intermittent or continuous in time. In an intermittent method, observations can be processed in small batches, which is usually technically convenient. In a continuous method, observation batches over longer periods are considered, and the correction to the analysed state is smooth in time, which is physically more realistic.

1.1.3 The data assimilation method

1.2 Advantages of Data Assimilation

Practically, data assimilation usually refers to the use of available measurements to correct a model's first prediction in space and time. It allows the optimal combination of model and observations to give improved forecasts and maximize the benefit from the observations. Here are some other advantages of data assimilation: **Improved Model Accuracy:** Data assimilation helps improve the accuracy of numerical models by incorporating real-world observational data. This is particularly important in situations where models alone may not accurately represent complex and dynamic systems. **Reduced Uncertainty:** By assimilating observational data into models, uncertainties in the initial conditions and model parameters can

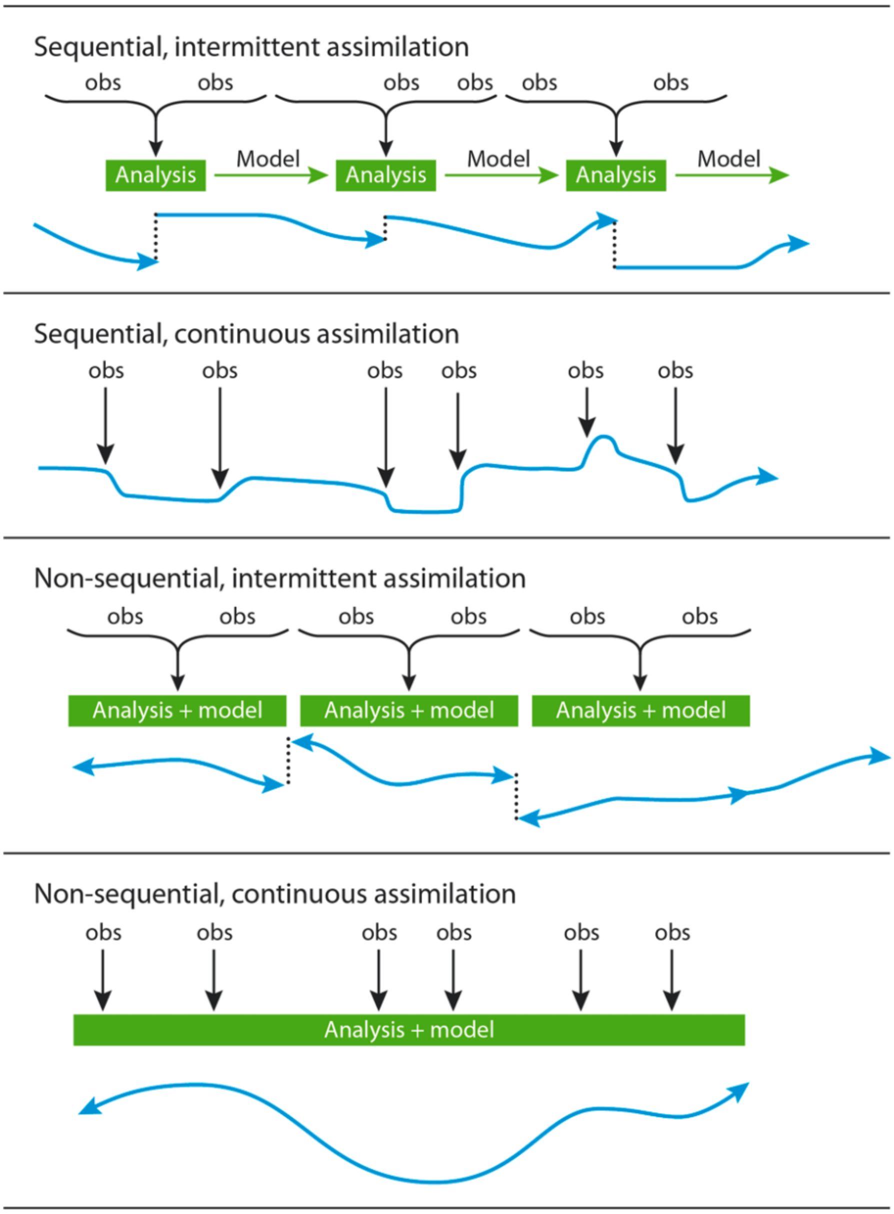


Figure 1: Representation of four basic strategies for data assimilation, as a function of time.

be reduced. This leads to more reliable and precise predictions, which is crucial in applications such as weather forecasting and climate modeling. **Adaptability to Changing Conditions:** Systems in the real world are dynamic and subject to change. Data assimilation allows models to adapt to evolving conditions by continuously incorporating new observational data, ensuring that predictions remain relevant and accurate over time. **Real-Time Applications:** Data assimilation enables real-time updating of model predictions based on the latest available observational data. This is essential for applications such as weather forecasting, where timely and accurate information is crucial. **Scientific Research and Discovery:** Data assimilation contributes to scientific understanding by providing a means to test and refine models against observational data. This iterative process can lead to new insights and discoveries about the behavior of complex systems.

1.3 Data Assimilation and AI Modeling

The integration of data assimilation and AI modeling can lead to more effective and accurate representations of complex systems. Both approaches bring unique strengths to the table, and their combination holds great potential for advancing scientific understanding and improving predictive capabilities across various disciplines.

2 Our Approach

2.1 Data Assimilation Methods

Before introducing the image below, it is essential to note that data assimilation methods are crucial in enhancing the accuracy of predictions across various scientific and technical disciplines. These methods are generally divided into two main categories: sequential and non-sequential. The sequential method includes techniques such as the Kalman Filter and optimal interpolation, which are applied iteratively as new data becomes available. On the other hand, non-sequential methods, such as 3DVAR and 4DVAR, process data in a single batch over a given time period.

After reviewing various data assimilation methods, we have chosen to use the Kalman Filter, a sequential method, for our project. This decision was made due to the Kalman Filter's ability to provide accurate estimates of the state of dynamic systems in real time, which is vital for our goal of precise tracking and forecasting.

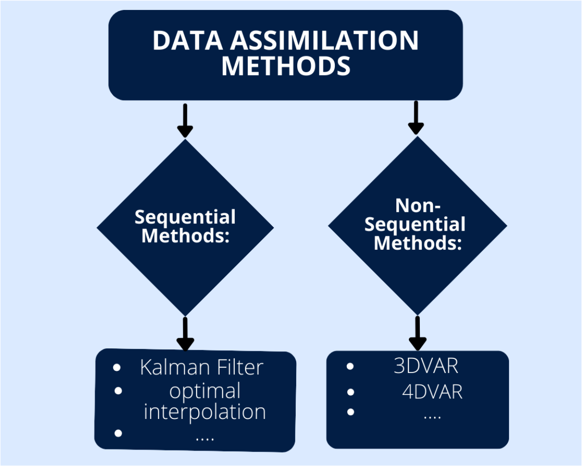


Figure 2: Data assimilation methods

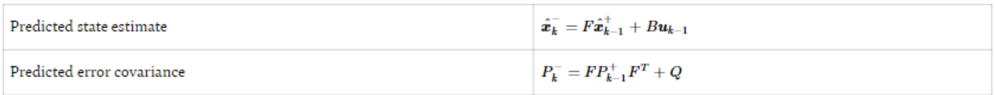
2.2 Kalman Filter Algorithm

The Kalman Filter is a pivotal estimation tool utilized extensively across various disciplines for its robustness in uncertain environments. This algorithm excels in deducing the concealed states of a dynamic system, adeptly handling measurements that may be marred by inaccuracies or ambiguity. It not only interprets the present state with precision but also projects future states with a remarkable degree of reliability, leveraging historical data and prior estimations.

The filter is named after Rudolf E. Kálmán (May 19, 1930 – July 2, 2016). In 1960, Kálmán published his famous paper describing a recursive solution to the discrete-data linear filtering problem.

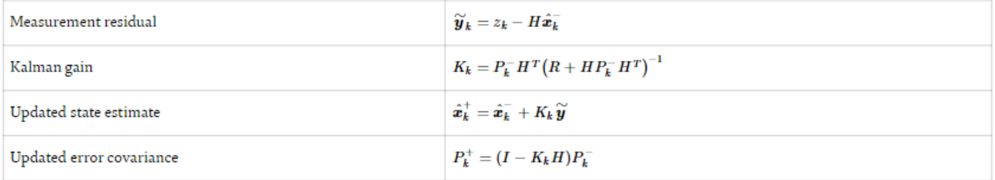
The Kalman Filter operates fundamentally in two stages: **the prediction phase** and **the update phase**. It's interesting to note that within the breadth of scholarly articles, these stages are sometimes alternatively termed "propagation" for prediction, and "correction" for the update. The Kalman Filter algorithm can be summarized as follows:

Prediction:



Predicted state estimate	$\hat{\boldsymbol{x}}_k^- = \boldsymbol{F}\hat{\boldsymbol{x}}_{k-1}^+ + \boldsymbol{B}\boldsymbol{u}_{k-1}$
Predicted error covariance	$\boldsymbol{P}_k^- = \boldsymbol{F}\boldsymbol{P}_{k-1}^+\boldsymbol{F}^T + \boldsymbol{Q}$

Update :



Measurement residual	$\widetilde{\boldsymbol{y}}_k = z_k - H\hat{\boldsymbol{x}}_k^-$
Kalman gain	$K_k = P_k^- H^T (R + HP_k^- H^T)^{-1}$
Updated state estimate	$\hat{\boldsymbol{x}}_k^+ = \hat{\boldsymbol{x}}_k^- + K_k \widetilde{\boldsymbol{y}}$
Updated error covariance	$P_k^+ = (I - K_k H) P_k^-$

In the above equations, the hat operator, , means an estimate of a variable. That is, x is an estimate of x . The superscripts – and + denote predicted (prior) and updated (posterior) estimates, respectively.

Kalman filters are predicated on the premise that both the process and measurement models adhere to linearity, meaning they can be characterized through matrices F, B, and H. Additionally, it assumes that the noise present in both the process and measurement is Gaussian and additive. Therefore, the optimality of the estimates yielded by a Kalman filter is contingent upon the fulfillment of these assumptions.

3 Developpement and Application

3.1 AI Model

For our project, we utilized an existing AI model. The AI model is a sophisticated neural network built within the TensorFlow environment, designed for predictive simulations in the realm of flotation cell operations. It harnesses the power of dynamic data collected from sensors, static user-configured flotation cell data, and unchanging ore kinetics data related to mineral composition.

The input parameters comprise a triad of crucial information:

Processing Feed Stream Data (Dynamic): Real-time data collected from sensors, providing a dynamic understanding of the operational conditions.

Flotation Cell's Data (Static): Configurable static data that users set before initiating predictions or simulations. It is unique to each flotation cell.

Ore Kinetics Data (Static): Unchanging data associated with the ore composition, which remains constant throughout the entire circuit.

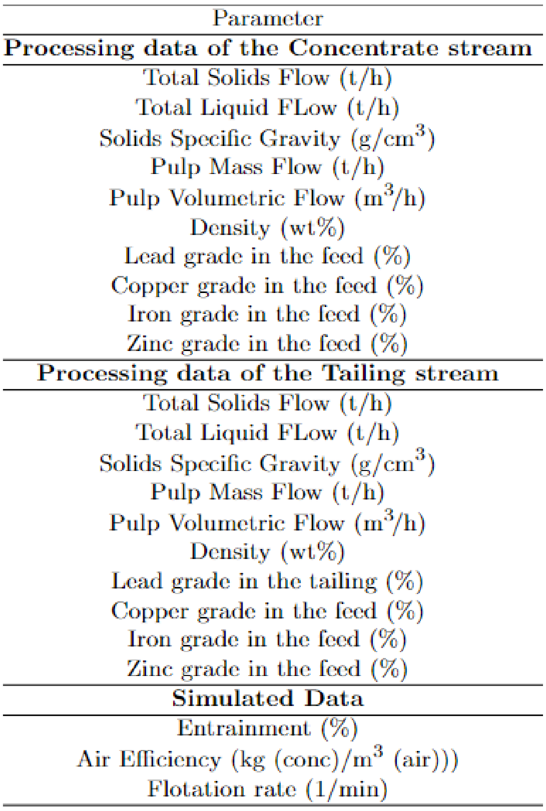
It's noteworthy that while Flotation Cell's Data varies for each cell, the Ore Kinetics Data remains consistent across the entire circuit.

The neural network processes input arrays and produces corresponding output arrays, aligning with the parameters of the flotation cell.

Based on these inputs, the model generates a set of 23 outputs

Processing data of the feed stream
Total Solids Flow (t/h)
Total Liquid FLow (t/h)
Solids Specific Gravity (g/cm^3)
Pulp Mass Flow (t/h)
Pulp Volumetric Flow (m^3/h)
Density (wt%)
Lead grade in the feed (%)
Copper grade in the feed (%)
Iron grade in the feed (%)
Zinc grade in the feed (%)
Flotation Cell's data
Cell's Volume (m^3)
Pulp Area (m^2)
Froth Thickness (mm)
Air Flow rate within the cell (m^3/min)
Ore Kinetics
R_inf_$CuFeS_2$ (%)
K_max_$CuFeS_2$ (1/min)
R_inf_PbS (%)
K_max_PbS (1/min)
R_inf_$Fe_{1-x}S$ (%)
K_max_$Fe_{1-x}S$ (1/min)
R_inf_ZnS (%)
K_max_ZnS (1/min)

Figure 3: Inputs



Parameter
Processing data of the Concentrate stream
Total Solids Flow (t/h)
Total Liquid FLow (t/h)
Solids Specific Gravity (g/cm^3)
Pulp Mass Flow (t/h)
Pulp Volumetric Flow (m^3/h)
Density (wt%)
Lead grade in the feed (%)
Copper grade in the feed (%)
Iron grade in the feed (%)
Zinc grade in the feed (%)
Processing data of the Tailing stream
Total Solids Flow (t/h)
Total Liquid FLow (t/h)
Solids Specific Gravity (g/cm^3)
Pulp Mass Flow (t/h)
Pulp Volumetric Flow (m^3/h)
Density (wt%)
Lead grade in the tailing (%)
Copper grade in the feed (%)
Iron grade in the feed (%)
Zinc grade in the feed (%)
Simulated Data
Entrainment (%)
Air Efficiency (kg (conc)/m^3 (air)))
Flotation rate (1/min)

Figure 4: Outputs

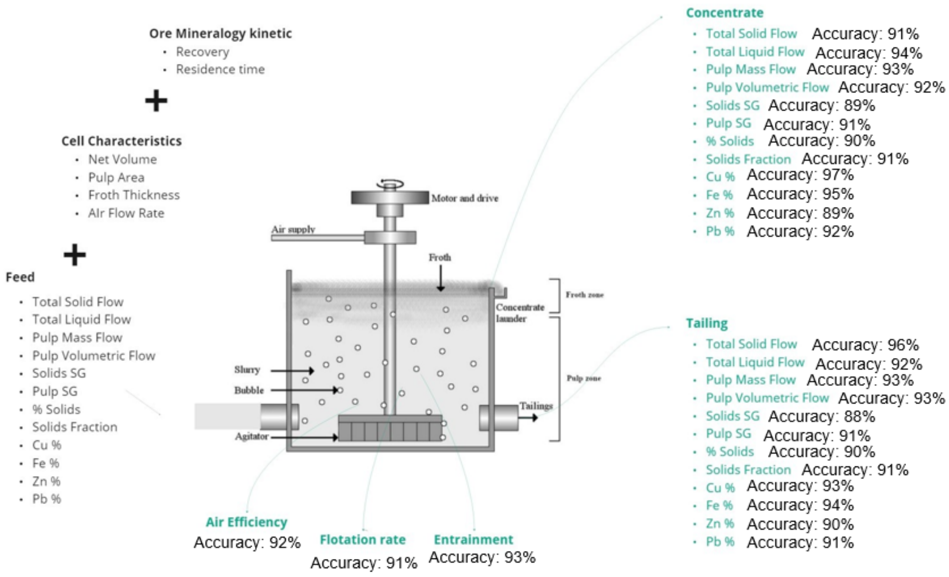
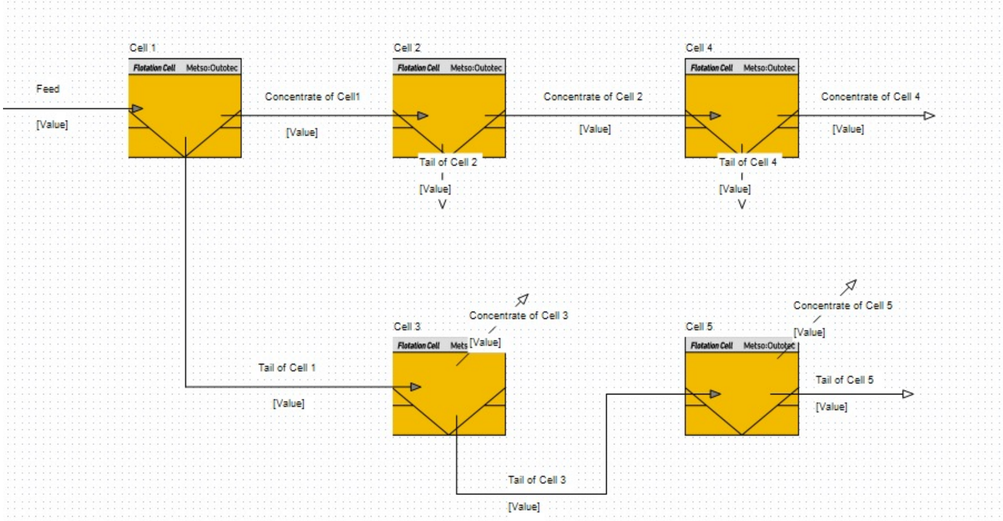


Figure 5: Illustration of the AI model

3.2 Simulation

In the context of our research, we meticulously followed a simulation encompassing five distinct cells, which we will refer to as Cell 1, Cell 2, Cell 3, Cell 4, and Cell5. Our objective was to analyze how predictions from a specific model behaved as we progressed from one cell to another.



At each stage of the simulation, we applied the AI model to generate predictions. Our approach aimed to comprehend how the quality of these predictions

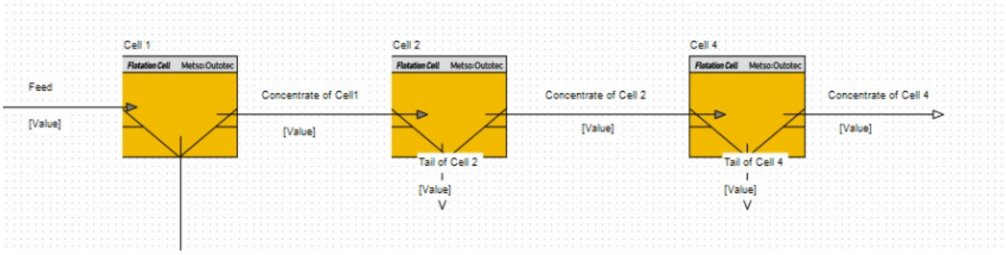


Figure 6: Integration of Data Assimilation using Kalman Filter for cells 1,2 and 4

evolved as we moved from one cell to another. We closely examined the model's performance and took note of any signs of degradation or improvement in its predictive capabilities. This analysis allowed us to gain a better understanding of the model's stability and effectiveness in different cellular contexts.

3.3 Integration of Data Assimilation using Kalman Filter

In this process, the Kalman Filter (KF) is utilized specifically for cells 1, 2, and 4. Initially, the input for cell 1 is derived from direct observations. This observational data is also fed into the Kalman Filter. Subsequently, the KF works to enhance the accuracy of the model's output for cell 1. While the improvement in accuracy is modest, it is still significant. After applying the KF to cell 1's data, the refined output is saved. This enhanced output, referred to as 'concentrated features', is then used as the input for cell 2. Following a similar procedure, the same approach is applied to cell 4.

3.4 Model Performance Enhancement

In our research, we focused on refining an existing AI model which initially demonstrated a commendable predictive accuracy with an R2 score of 0.9125. Our objective was to elevate its predictive performance, specifically targeting an improvement in the R2 score.

3.4.1 Methodology

Data Preprocessing: The model's input data, underwent standardization. This preprocessing step was crucial for maintaining consistency in the scale of the input features, achieved through the StandardScaler from scikit-learn.

Model Architecture Review: The existing model was a sophisticated neural network developed using TensorFlow. We identified that by fine-tuning the model's architecture and hyperparameters, we could potentially unlock higher levels of accuracy.

Hyperparameter Optimization Strategy: Our approach was centered around optimizing key hyperparameters: optimizer choice, number of neurons, dropout rates, and regularization parameters (L1 and L2). The selection of these parameters was critical in balancing the model's ability to learn from the data against the risk of overfitting.

3.4.2 Implementation

Customized Hyperparameter Tuning Loop: Instead of relying on conventional grid search techniques, we implemented a custom tuning loop. This allowed for a meticulous exploration of the hyperparameter space, where each combination was rigorously tested for its impact on the R2 score.

Early Stopping Integration: To circumvent overfitting during training, we employed an early stopping mechanism. This process halted training when the model's performance on validation data plateaued, ensuring the model's generalizability.

Performance Evaluation Metric: The R2 score was employed as the primary metric for assessing the model's predictive accuracy. This metric facilitated a direct comparison of the enhanced model's performance against its initial state.

3.4.3 Results

Achieved Performance Uplift: Post enhancement, the model achieved an R2 score of 0.9343, marking a significant improvement from its previous state. This increment not only signifies the success of our optimization techniques but also highlights the model's increased reliability in predictive scenarios.

Optimal Hyperparameter Configuration: The optimal model configuration utilized the Adam optimizer, with each hidden layer containing 64 neurons. A dropout rate of 10 Percent and regularization rates (L1 and L2) of 0.001 were found to be most effective in achieving this superior performance.

3.4.4 Discussion

Significance of Results: The uplift in the R2 score to 0.9343 is indicative of the model's enhanced accuracy. This advancement underscores the value of systematic hyperparameter tuning in AI model refinement.

Future Research Directions: While the current improvements are substantial, exploring automated machine learning (AutoML) for hyperparameter optimization and testing the model in real-world scenarios could be potential avenues for future work. Continuous monitoring and iterative refinements will remain pivotal in maintaining the model's efficacy.

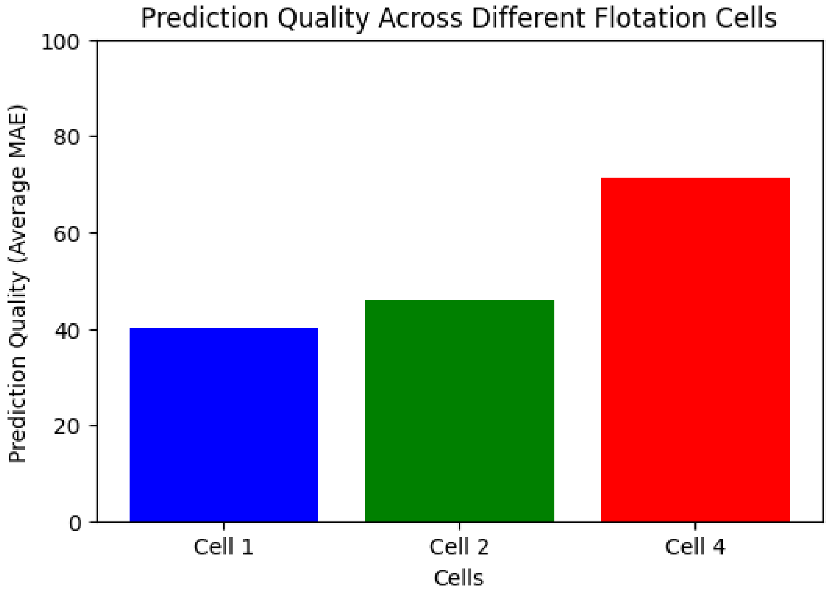
3.4.5 Conclusion

The systematic enhancement of the existing AI model has led to a noteworthy increase in its predictive accuracy. Through careful hyperparameter tuning and architectural adjustments, we have elevated the model's R2 score, thereby extending its applicability and robustness. This achievement not only contributes to the model's practical utility but also adds to the growing body of knowledge in neural network optimization.

4 Results and Discussions

4.1 Comparative Analysis of Model Predictive Accuracy Across Flotation Cells

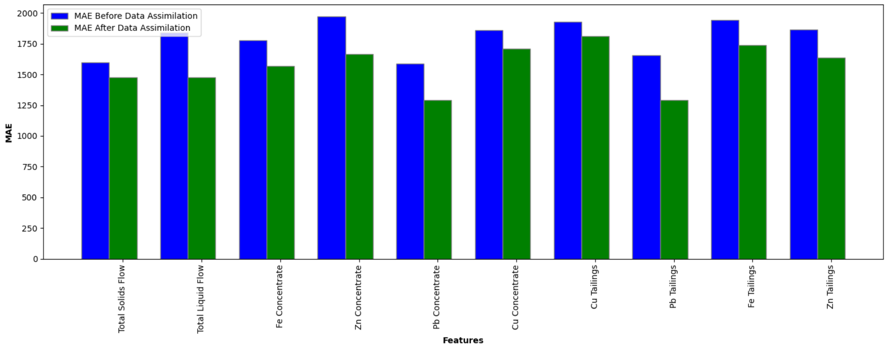
Let's examine the following bar graph to understand how the predictive accuracy of our model varies across the three flotation cells, as indicated by the Mean Absolute Error (MAE) for each cell



The bar graph above illustrates the average Mean Absolute Error (MAE) for three different flotation cells. As depicted, there is a noticeable increase in the MAE from Cell 1 to Cell 4.

This means that the predictive accuracy of the AI model progressively decreases from Cell 1 to Cell 4.

4.2 Accuracy of The Model Before and After Kalman Filter



Here's the bar plot showing the Mean Absolute Error (MAE) scores before and after data assimilation for the various features. As shown, the application of the data assimilation process, in this case using a Kalman filter, appears to reduce the MAE scores, indicating an improvement in the predictions.

5 Conclusion

This project has highlighted the effectiveness of data assimilation, through the application of the Kalman filter algorithm, in improving artificial intelligence (AI) models. Our results demonstrate a significant improvement in prediction accuracy, marked by a reduction in the Mean Absolute Error (MAE) and an increase in the coefficient of determination (R^2). This study reinforces the idea that data assimilation is crucial for the evolution of AI models, particularly in contexts requiring real-time data integration. The advancements made here open promising prospects for future research in the optimization of AI models and their application in various fields.

6 Acknowledgements

This project would not have been possible without the invaluable support and guidance of Mr. Oussama Hasidi. His expertise and insightful mentorship were pivotal to our success. We are deeply grateful for the dedication and substantial time he invested in guiding our team. In addition, we are immensely grateful to our mentor, Prof. El Hassan Abdelwahed. His teachings and steadfast support have been fundamental to our understanding and progress.