BX06 - Lithological Layer Prediction (LLP) through Machine Learning in Mineral Exploration at Paragominas Bauxite Province, Brazil

Acácio Nunes de Pina Neto[1], Jefferson Klister[2], Andras Fulop[3], Tamas Petz[3], Gustavo Loureiro[1], Dayane do Nascimento Coelho[1], Ricardo Radtke[1], Bruno Lima Gomes[4] and Helcio José Prazeres Filho[5]

1. Exploration Geologist
2. System Analyst
3. Solutions Architect, Datapao
3. Senior Data Scientist, Datapao
4. Mineral Exploration Manager
5. Mineral Exploration Senior Manager

Hydro Paragominas, Paragominas, Brazil

Corresponding author: acacio.pina.neto@hydro.com

Abstract

An accurate database is essential for geological interpretation, geological modeling, and decision-making in mineral exploration and mining activities. To ensure database reliability, validation of chemical samples is necessary to identify and correct inconsistencies as well as to minimize classification errors during data input. After a brief geological description, sampling, and lithological classification, a final validation based on laboratory results is required. Which can be a tedious and time-consuming process consisting of a set of conditions and filters applied using Excel to validate initial classification and identify outliers. This paper investigates a means to improve this process by applying a Deep Neural Network (DNN) classifier running in Azure Databricks environment (a Microsoft® platform) to make lithology predictions and database validation from chemical samples sourced from two different targets located in the Paragominas Bauxite Province (PBP), northeastern Pará State (Brazil). The application of Lithological Layer Prediction (LLP) using machine learning algorithms was found to significantly improve database validation. By analyzing and interpreting a vast amount of geological data using machine learning techniques, the accuracy and speed of lithology prediction was significantly improved. This technology has proven to be a valuable tool in identifying and characterizing bauxite deposits, allowing for more efficient and targeted exploration efforts.

Keywords: Lithology prediction, Machine learning, Bauxite, Paragominas Bauxite Province.

1. Introduction

Robust review and database validation are essential to provide valuable inputs for subsequent geological modelling and mineral resources estimates, as they ensure accuracy, completeness, and consistency of the data used in the analysis. Complexities involving historical data, many drilling campaigns as well as copious amounts of time for manual data interpretation have unique challenges and can introduce potential errors. Without proper validation, the results of the analysis can be unreliable, and decisions made based on the analysis could be incorrect.

In standard mineral exploration procedures, after a brief geological description (logging), sampling, and lithological classification, a final validation based on laboratory results is required (Figure 1). Although the latter is highly necessary, it is a time-consuming process consisting of a set of conditions and filters applied using Excel to validate initial classification and identify inconsistencies/outliers.

Figure 1. Mineral exploration workflow.

Due to these challenges, the prediction of lithological layers using machine learning (ML) models has been extensively explored [1][2][3][4]. While various machine learning techniques have been applied to lithological layer prediction, the utilization of deep learning (DL) models in conjunction with chemical components as features in the context of bauxite mines remains relatively new.

Many studies have incorporated traditional supervised techniques, such as ensemble tree models [5] or support vector machines (SVMs) [6] for lithological layer prediction [7]. Sebtosheikh *et al.* (2015) [8] demonstrated the effectiveness of SVMs in lithology prediction and provided insights on selecting optimal kernel functions and parameters for small datasets. Dev *et al*. (2019) [9] proposed the use of Extreme Gradient Boosting Trees for prediction using seismic logs, while Martin *et al.* (2021) [10] compared Extreme Gradient Boosting Trees with Convolutional Neural Networks [11] (CNNs) to improve prediction accuracy.

DL algorithms achieve high performance for classification, regression, clustering, and other applications [12][13][14]. The multiple hidden layers of DNN, the activation functions commonly employed, and methods of regularization to prevent overfitting, among other benefits allow DL algorithms to learn hierarchical feature representations of data with multiple levels of abstraction [13]. These methods create a powerful tool to help identify anomalies and reveal patterns in large datasets with less manual feature engineering than traditional ML methods.

2. Methodology

The description below presents a DNN-based method from chemical samples analysis for validation and lithological classification in geochemical data. This approach is entirely data-driven and, once the network is trained, delivers the results in real time by predicting the samples classification from input data in a single step.

2.1 Study Area

The study area is located at Paragominas Bauxite Province (PBP) central domain, in the northeast region of the state of Pará, in the Eastern Amazon, Brazil. The PBP represents one of the most important, extensive, and dense groupings of bauxite deposits in Brazil, with a potential of more than 3 billion tonnes of metallurgical ore, about 70% of Brazil's total bauxite reserves [15]. PBP is characterized by a relief of plateaus covered by a thick layer of clays (Belterra clay) and ferro-aluminous crusts. The formation of these deposits was originated by the lateritic alteration of siliciclastic deposits from the Cretaceous, in this case, sediments from the Itapecuru and the Ipixuna Formation, during the Paleogene [15] (Figure 2).

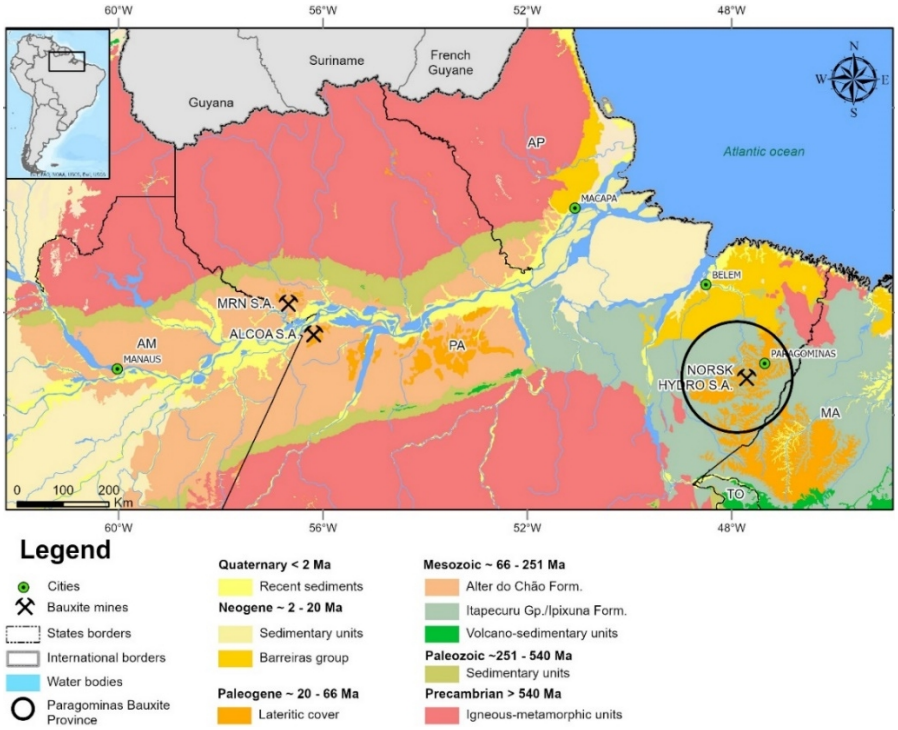

Figure 2. Simplified geological map of northern Brazil with the distribution of the main lithostratigraphic units.

The Paragominas Province central domain comprises many bauxite deposits including the Miltonia 5 (Mil5) and Miltonia 3 (Mil3) plateaus, where the Hydro Paragominas bauxite mine has operated since 2012. The mine is located in the municipality of Paragominas, 356 km from its capital, Belém do Pará.

The lateritic profile generally comprises five main lithotypes characterized by clear textural, compositional, and color differences and well-defined contacts (Figure 3).

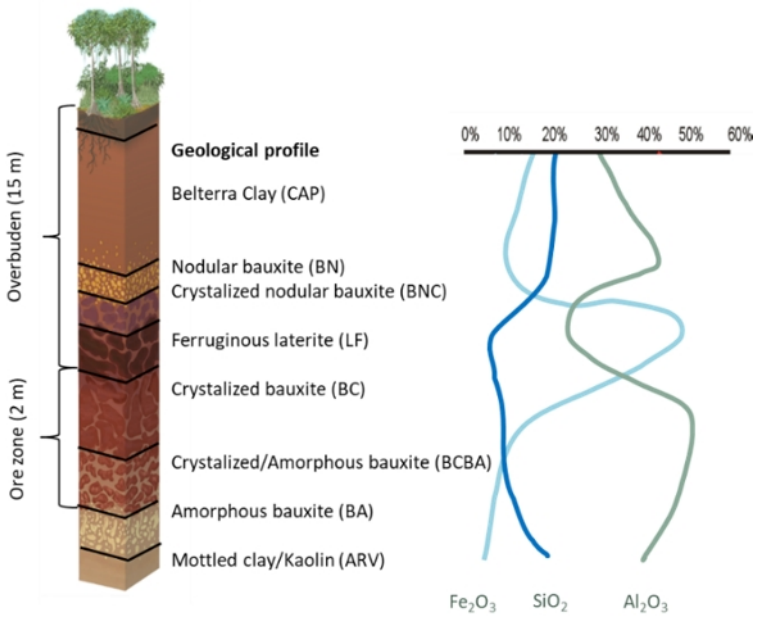

Figure 3. Regional geological profile.

A brief description of these units follows (from bottom to top): A) The first lithotype from base to top of lateritic profile is the bottom clay (ARV), which has a gradual transition from the fine-grained kaolinitic sandstone bedrock of Itapecuru/Ipixuna formations. This layer includes typical saprolitic and mottled zone ending in the main bauxite zone; B) Lower Bauxite horizon is formed of a massive bauxite layer of reddish color with abundant, millimeter-sized gibbsite crystals and iron oxides, it is subdivided in Crystallized Bauxite with "Amorphous" Bauxite (BCBA) and Crystallized Bauxite (BC); D) Ferruginous Laterite (LF) is composed of goethite and hematite pisolites in a massive texture; E) Nodular Bauxite (BN), being composed of heterogeneous gibbsite nodules, formed by amorphous bauxite immersed in a kaolinitic matrix; F) Belterra clay (CAP) , discordant and occurs with frequent ferruginous pisolites found among gibbsite nodules. This geological formation covers the lateritic profile and has 5 to 20 meters of thickness, being constituted by a homogeneous sequence of kaolinitic clays.

2.2 Input Data

The dataset consisted of a series of bauxite drill sample sequences. Each sequence was divided into a number of sub-samples, ranging from 25 to 75 cm long. These sub-samples included positional information, chemical analysis measurements, and the initial layer classification determined manually by geologists during the sampling process.

The positional features encompassed X-Y-Z coordinates, layer start-end positions, and length. The chemical analysis data included measurements of Available Alumina, Reactive Silica, Total Alumina, Iron oxide, Silicon dioxide, Titanium dioxide, as well as the loss on ignition (LOI). Additionally, two layer-related attributes were available for historical records: the initial layer classification provided by the geologist during sampling and the final (re)classification derived from the initial information and chemical analysis results. The final (re)classification served as our target variable (Table 1).

The lithological layers around the drilling sites were categorized into eight distinct layers (Figure 3). It is worth noting that there might be additional layers, but they were excluded from the samples since they represent layer name variations and/or historical subcategories.

Table 1. Initial features

FEATURE NAME	DESCRIPTION	DATA TYPE
DRILLHOLE	Drill sample sequence identifier	string
YEAR	Year of the drilling	integer
X	Sample X coordinate	double
Y	Sample Y coordinate	double
Z	Sample Z coordinate	double
FROM	Subsample start point (in cm)	double
TO	Subsample end point (in cm)	double
LENGTH	Subsample length (in cm)	double
Al_2O_3 Av.	Chemical analysis result: Available Alumina content	double
SiO_2 Reac.	Chemical analysis result: Reactive Silica content	double
Fe_2O_3	Chemical analysis result: Iron oxide content	double
Al_2O_3	Chemical analysis result: Total Alumina content	double
SiO2	Chemical analysis result: Silicon dioxide content	double
TiO2	Chemical analysis result: Titanium dioxide content	double
LOI	Chemical analysis result: Loss on ignition	double
REC	Chemical analysis result: Mass recovery	double
LITHODES	Initial subsample classification	string
LITHO_RECLASS	Final subsample classification, target value	string

2.3 Exploratory Analysis and Feature Selection

To gain more detailed insights into the features' usefulness, an exploratory analysis focusing on the chemical components of each layer was performed. The analysis revealed several important findings: a) many chemical components appeared in small quantities, b) most components

exhibited minimal variation between layers, and c) the layer compositions overlapped, making them unsuitable for accurate prediction.

Based on these findings, we selected the three most impactful chemical species, namely: Available Alumina (Al_2O_3 Av.), Reactive Silica (SiO_2 reat.), and Iron Oxide (Fe_2O_3). To better represent their relationships, we generated specific derived features, including their ratios, sum of values, and relative percentages against their sum. For convenience, we encoded the layers as numbers and introduced the "depth" feature to represent the order of each subsample within the drill sample. Finally, we added the previous layer's index as an extra feature. Further details can be found in Table 2.

Table 2. Derived features and their calculation.

DERIVED FEATURE CALCULATION	FEATURE NAME	DESCRIPTION
Al/Si	Al/Si	aluminium - silicon ratio
Fe_2O_3/Si	Fe/Si	iron oxide - silicon ratio
Al/Fe_2O_3	Al/Fe	aluminium - iron oxide ratio
$Al + Si + Fe_2O_3$	important	sum of the "important" features
$Al/(Al + Si + Fe_2O_3)$	Al_ratio	aluminium - important ratio
$Fe_2O_3/(Al + Si + Fe_2O_3)$	Fe_ratio	iron oxide - important ratio
$Si/(Al + Si + Fe_2O_3)$	Si_ratio	silicon - important ratio
Map lithodes to numbers	lithode_order	Lithology order number
Map (re)classified lithodes to numbers	lithode_reclass_order	(Re)classified lithology order number
Subsample index starting from the top	depth	layer position within the drill sample
Previous subsample's layer index	prev_layer	the previous layer's index

Following the feature generation, we examined the Pearson correlations [16] between the original and derived features, as well as the target variable (Figure 4). Given the limited size of our historical data, we aimed to eliminate redundant and unimportant features. The elimination process considered different feature groups: metadata features (drill identifier and year of drilling), positional data (coordinates and length information), and chemical analysis columns. The elimination was performed section-by-section.

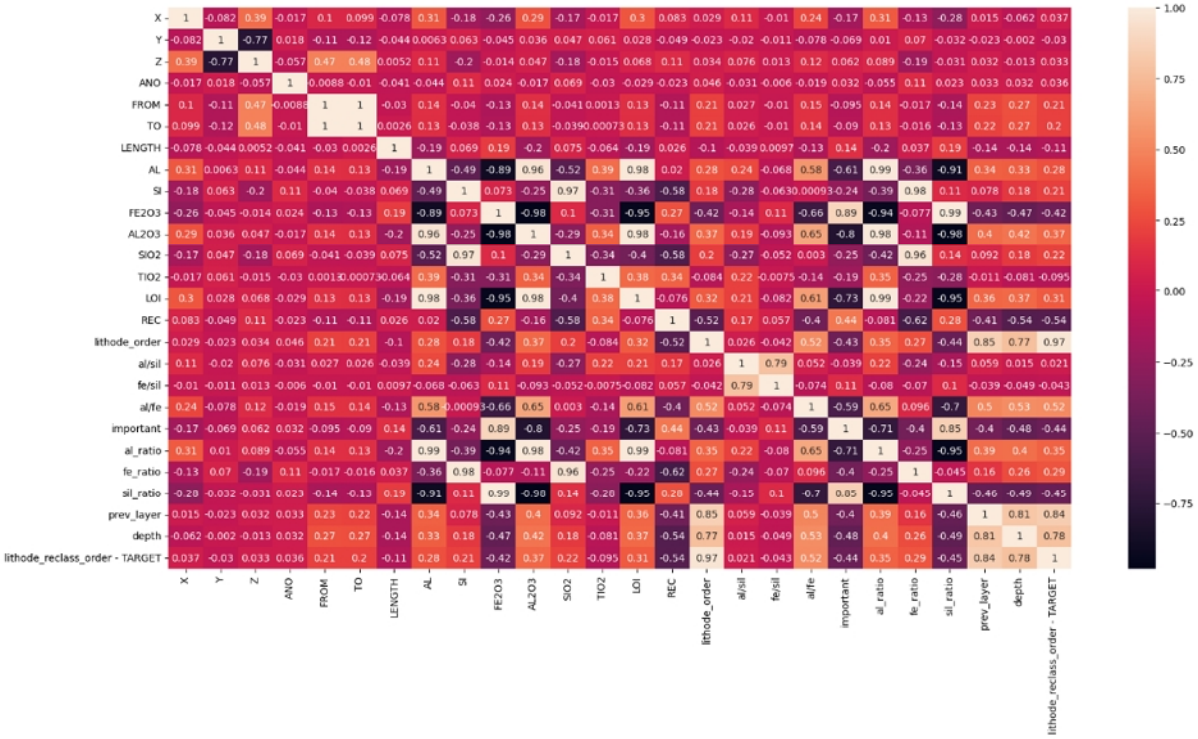

Figure 4. Pearson correlation heatmap between the features and target.

Among the metadata columns, only the drill identifier was retained to ensure the ability to handle samples as sequences, as further discussed in the modeling section.

The X, Y, Z coordinates were removed, as they did not provide useful information relating to the layer sequences. Although the FROM, TO, and LENGTH features initially seemed important, after completion of the initial modeling, their contribution was considered negligible. Notably, we discovered that the previous layer and the initial prediction were strong indicators. The inclusion of these features posed a significant challenge during modeling and greatly influenced our evaluation approach.

From the main chemical analysis features, this research focused on the derived features. Among the seven related features, we narrowed down the selection to the top three, namely the relative ratios of different chemical components against their sum (Al_ratio, Fe_ratio, Si_ratio).

The final set of features can be found in Table 3.

Table 3. Final selection of features for modeling.

FEATURE NAME	DESCRIPTION	DATA TYPE
Drillhole	Drill sample sequence identifier	string
Al_ratio	Aluminium - important ratio	double
Fe_ratio	Iron oxide - important ratio	double
Si_ratio	Silicon - important ratio	double
lithode_order	Lithology order number	integer
depth	Layer position within the drill sample	integer
prev_layer	the previous layer's index	integer
lithode_reclass_order	(Re)classified lithology order number (target)	integer

2.4 Modelling

Challenges

The primary challenge is not to find an accurate layer-by-layer classifier: even simple tree methods can achieve satisfactory performance. Instead, the challenge is to adhere to the rising constraint on the lithological layer orders within the drill samples. To address this challenge, our modeling approach began by training simple models on the Mil3 and Mil5 sites separately, as well as on their combined dataset. During this initial training, we treated each sub-sample as an individual point and captured the layer relationships using the 'prev_layer' feature, which encodes the sequential nature of the lithological layers.

We employed a Decision Tree (DT) and a Random Forest (RF) model using the scikit-learn module [17] (version 1.1.1), and an Extreme Boosted Tree model using the XGBoost library [18] (version 1.7.5). These base models provided encouraging results (Table 4), based on their classification reports. For the model's performance evaluation, metrics such as Accuracy, Recall, and F1-score were used. Accuracy represents the number of correctly classified data instances over the total number of data instances. Precision tells us what percentage of all the Positive predictions made by the model were accurate. Recall is also known as 'sensitivity' or 'true positive rate' and relates to what percentage of all the actual Positives were accurately predicted by the model, while F1 score is the harmonic mean of precision and 'recall' and is a better measure than accuracy.

It is important to note that the base model's high performance is partly due to the inclusion of the initial lithology classification (LITHODES) as a feature. The classification target only differs from the initial classification approximately 8 % of the time. Without this feature, the performance of these models decreases by roughly 30 %. Another issue is that these models often disregard lithostratigraphy, in which layers' positions can be inverted.

In summary, the main challenges were as follows:

- The predicted layer orders must form a non-decreasing sequence within a drill sample.
- The prediction relies on the initial layer classification as a feature.
- The model needs to validate or reclassify the initial classification.
- Different mines may exhibit significant variations.

Table 4. Classification report of the baseline models.

Decision Tree

layer name	mil3				mil5				generic			
	precision	recall	f1-score	support	precision	recall	f1-score	support	precision	recall	f1-score	support
CAP	0.00	0%	0%	2	0%	0%	0%	2	0%	0%	0%	0
BN	0.97	97%	97%	853	89%	85%	87%	752	92%	92%	92%	1702
BNC	0.96	96%	96%	441	82%	86%	84%	513	90%	91%	90%	1322
LF	0.97	97%	97%	922	83%	77%	80%	168	93%	93%	93%	628
BC	0.98	98%	98%	1859	93%	94%	93%	532	96%	96%	96%	2384
BCBA	94%	93%	94%	1241	89%	89%	89%	457	92%	91%	91%	1736
BA	85%	87%	86%	537	58%	65%	61%	211	74%	74%	74%	780
ARV	88%	90%	89%	263	76%	74%	75%	300	76%	78%	77%	615
accuracy			95%	6118				0.84			90%	9169
macro avg	82%	82%	82%	6118	71%	71%	71%	2935	77%	77%	77%	9169
weighted avg	95%	95%	95%	6118	85%	84%	84%	2935	90%	90%	90%	9169

Random forest

layer name	mil3				mil5				generic			
	precision	recall	f1-score	support	precision	recall	f1-score	support	precision	recall	f1-score	support
CAP	0%	0%	0%	2	0%	0%	0%	2	0%	0%	0%	2
BN	98%	98%	98%	853	86%	88%	87%	513	94%	94%	94%	1702
BNC	98%	97%	97%	441	91%	83%	87%	168	93%	93%	93%	1322
LF	98%	98%	98%	922	92%	88%	90%	752	97%	95%	96%	628
BC	99%	98%	99%	1859	97%	95%	96%	532	98%	97%	98%	2384
BCBA	96%	95%	95%	1241	92%	94%	93%	457	96%	94%	95%	1736
BA	89%	91%	90%	537	70%	64%	67%	211	81%	88%	84%	780
ARV	88%	92%	90%	263	66%	79%	72%	300	77%	79%	78%	615
accuracy			97%	6118			87%	2935			93%	9169
macro avg	83%	84%	83%	6118	74%	74%	74%	2935	80%	80%	80%	9169
weighted avg	97%	97%	97%	6118	88%	87%	87%	2935	93%	93%	93%	9169

XGBoost mil3

layer name	mil3				mil5				generic			
	precision	recall	f1-score	support	precision	recall	f1-score	support	precision	recall	f1-score	support
CAP	0%	0%	0%	2	0%	0%	0%	2	0%	0%	0%	2
BN	98%	98%	98%	922	93%	88%	91%	752	96%	93%	95%	1702
BNC	98%	98%	98%	853	86%	88%	87%	513	93%	95%	94%	1322
LF	98%	96%	97%	441	91%	80%	85%	168	97%	94%	95%	628
BC	99%	98%	99%	1859	97%	95%	96%	532	98%	97%	98%	2384
BCBA	96%	95%	96%	1241	93%	93%	93%	457	96%	94%	95%	1736
BA	90%	92%	91%	537	71%	63%	67%	211	80%	89%	84%	780
ARV	83%	92%	87%	263	64%	82%	72%	300	79%	78%	78%	615
accuracy			96%	6118			87%	2935			93%	9169
macro avg	83%	84%	83%	6118	74%	74%	74%	2935	80%	80%	80%	9169
weighted avg	97%	96%	96%	6118	88%	87%	87%	2935	94%	93%	93%	9169

To address these challenges, we explored several alternatives to improve the performance of the baseline models:

- Building separate models for different mines: This approach proved effective in enhancing overall performance.
- Using boosted trees instead of 'bagging' [19]: While boosting trees considerably improved performance, it still could not handle the non-decreasing layer order constraint.
- Changing the target from layer order to a binary classification problem (validate/reclassify): this modification did not yield improved performance.
- Considering the drill sample as a sequence of layer orders and attempting to predict a complete sequence in a single prediction: This approach successfully resolved the non-decreasing constraint and provided satisfactory results in terms of validation and reclassification.

Our proposed model

Considering observations made during the initial modeling phase and drawing inspiration from related works [8][10], we decided to leverage the sequential nature [20] of the lithological layers. Our proposed model consists of a neural network with a single layer of Long-Short Term Memory [21] (LSTM) units followed by a *time distributed* dense layer.

Given the varied lengths of the drill samples, we standardized the sequence lengths using 'padding and cutting' techniques [22]. These functions come from the need to encode sequence data into contiguous batches. To make all sequences in a batch fit a given standard length, it is necessary to pad or truncate some sequences. Two types of padding were employed: a generic padding (PAD) to indicate empty items at the beginning of the sample sequence, and a "start" signal padding (CLS) to indicate the start of the sequence.

For preprocessing, we encoded the layer names as numbers, applied min-max scaling to the features, and one-hot encoded the target variable. The subsamples of the drill samples were then concatenated into matrices of size `m` x `n`, where `m` representing the number of used features, and `n` the maximum number of subsamples within the drill samples in the dataset. The LSTM layer encoded the sequential information within the samples, while the dense layer facilitated the multiclass classification. Figure 5 illustrates the model's architecture.

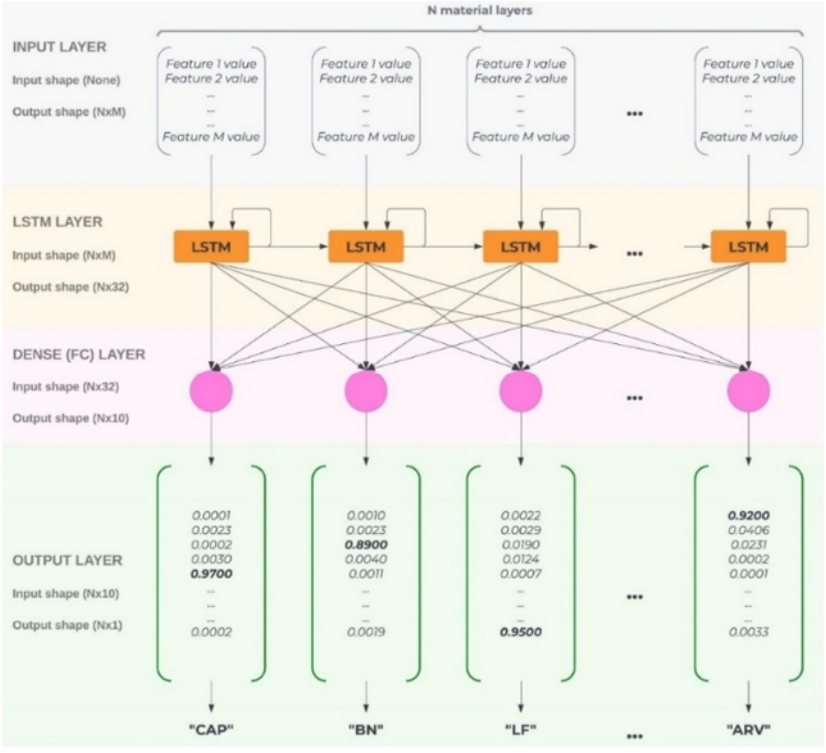

Figure 5. Model architecture.

One significant improvement during model development was the use of two types of padding. This modification allowed the model to discard the empty padding at the beginning of shorter samples, while maintaining performance. Another key enhancement was the consideration of the lithological layers as being in an ordinal scale. These were encoded accordingly when being used as features, specifically the lithological_order and prev_layer features. This decision served two purposes: first, it allowed us to express that the lithological layers form a non-decreasing sequence; second, since lithological layers do not have sharp boundaries, the cutting of drilling samples may not be ideal, potentially resulting in subsamples containing parts of two neighboring

layers. By employing the ordinal scale, we can indicate that layers closer to each other exhibit more similar chemical compositions.

We recognized that different mining sites exhibit distinct characteristics. Therefore, we made the decision to train separate models for each site, in addition to a generic model that can handle previously unseen sites. This approach not only improved the model's performance on specific sites, but also allowed other mining sites to test our solution.

The trained models are packed into an MLflow model [23], and the underlying models are selected dynamically during prediction time. If there is no specific model trained for the given drill sample, the generic model is used.

3. Results

3.1 Performance

In our evaluation of the proposed model, we focused on improving the recall metric, which measures the model's ability to detect cases where the final classification differs from the initial classification. As shown in Table 5, the classification metrics of the proposed model outperformed those of the baseline models. This improvement is particularly noticeable in the case of the Mil5 mine site, which has a smaller training dataset and similar lithological layers. Our model achieves a recall of 0.94, compared to 0.87 in the case of the Mil5 mine.

Table 5. Classification report of our suggested model.

Model	M3 model				M5 model				Generic model			
	precision	recall	f1-score	support	precision	recall	f1-score	support	precision	recall	f1-score	support
CAP	0%	0%	0%	1	100%	100%	100%	341	0%	0%	0%	4
BN	96%	99%	98%	790	91%	95%	93%	491	91%	95%	93%	1292
BNC	97%	96%	97%	436	88%	87%	88%	167	92%	90%	91%	564
LF	99%	97%	98%	977	100%	100%	100%	2498	95%	94%	94%	1674
BC	98%	99%	99%	1776	92%	93%	93%	582	98%	97%	98%	2348
BCBA	96%	97%	96%	1182	84%	88%	86%	465	94%	95%	95%	1622
BA	93%	92%	92%	591	65%	56%	60%	217	84%	87%	86%	811
ARV	97%	88%	92%	255	85%	80%	82%	314	89%	84%	86%	528
CLS	100%	100%	100%	753	94%	94%	94%	722	100%	100%	100%	1093
PAD	100%	100%	100%	6040					100%	100%	100%	8645
accuracy			99%	12801			94%	5797			97%	18581
macro avg	88%	87%	87%	12801	89%	88%	88%	5797	84%	84%	84%	18581
weighted avg	99%	99%	99%	12801	94%	94%	94%	5797	97%	97%	97%	18581

While classification metrics are important for evaluation, the true difference in performance can be observed in the rate of errors related to the rising constraint (layers' positions inversion). Table 6 illustrates this comparison, where our model significantly outperforms the baseline models by producing at least 3.5 times fewer errors at the subsample level (up to 4.2 times) and at least 1.7 times fewer errors at the drill sample level (up to 3.4 times).

The difference in the test data size between our proposed solution and the baseline models is that the baseline models could not handle missing values, whereas our model could. In the generic use-case, the proposed model achieved a layer count of 68 (0.74 % of total layers) and a drill sample count of 48 (4.18 % of total drill samples), indicating superior performance in adhering to the rising constraint. For comparison, the baseline models, including Decision Tree, Random Forest, and XGBoost Classifier, achieved layer counts ranging from 246 to 282 (2.68 % to 3.09 % of total layers) and drill sample counts ranging from 82 to 162 (7.14 % to 14.11 % of total drill samples).

Table 6. Comparison of the non-decreasing constraint errors.

MODEL	Nº LAYER	% LAYERS	Nº DRILL SAMPLE	% DRILL SAMPLE
LSTM	68	0.74%	48	4.18%
Number of Drill Samples	9176	100.00%	1149	100.00%
MODEL	**Nº LAYER**	**% LAYERS**	**Nº DRILL SAMPLE**	**% DRILL SAMPLE**
Decision Tree	246	2.68%	162	14.11%
Random Forest	282	3.09%	98	8.54%
XGBoost Classifier	271	2.96%	82	7.14%
Number of Drill Samples	9169	100.00%	1148	100.00%

We also examined the type of prediction errors, as shown in the confusion matrices in Figure 6. These matrices highlight the differences in the difficulty of lithological layer prediction between our example mines. Specifically, the Mil5 mine shows that distinguishing the transitional layers BC, BCBA, and BA is challenging due to their similar chemical composition. Furthermore, there appears to be a similarity between LF and BN layers, as they often appear at the top of the drill samples.

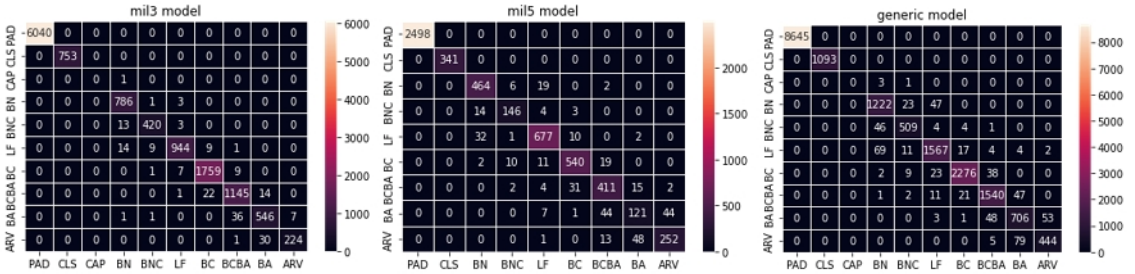

Figure 6. Confusion matrices of the trained models.

3.2 Software Ecosystem

In addition to achieving a model performance that is comparable to human-level classification, we have developed a sophisticated workflow capable of generating predictions using an uploaded Excel export of the chemical analysis results. By running this workflow, the uploaded file is enriched with the prediction results. The prediction runtime, starting from a cold state, was approximately 8 minutes (including the cluster startup time), while running inference on the entire historical dataset took 7 minutes and 52 seconds (Figure 7).

Retraining the model is a straightforward process that involves linking the historical chemical analysis and classifications, followed by executing the training pipeline. The system is scalable and easily extendable to accommodate new mine sites. The complete training process takes approximately 15 minutes to complete.

The prediction process provides additional features alongside the predictions. One such feature is the "fixed prediction" column, which addresses non-decreasing constraint violations by applying heuristics. The fixed records are marked to indicate the correction. Another feature is the confidence value associated with each prediction, which enables geologists to identify cases that pose more significant challenges for classification and may require manual review.

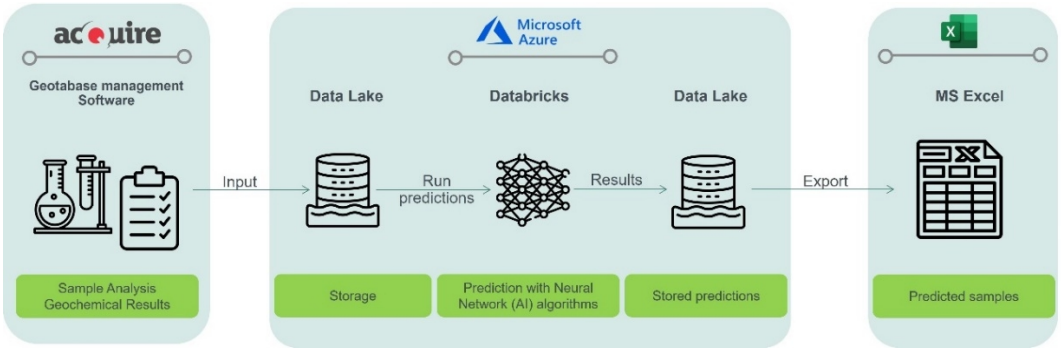

Figure 7. Solution workflow.

Moreover, the inference process generates a prediction report that includes reclassification metrics, capturing the original lithological layer and the reclassified lithological layer with corresponding layer counts. The report also presents an analysis of prediction confidence values, counts of uncertain predictions, and identifies order errors, both within the original classification and in the prediction.

3.3 Further Work

While the system developed is robust and straightforward to use, it does require some initial training and access to Azure Databricks. To enhance usability, we propose creating a convenient front-end application that can facilitate the entire prediction process, including file upload, workflow execution, and result download.

Also, the current solution utilizes a relatively shallow network. By gathering more historical data from other sites, it would be possible to train deeper networks, potentially enhancing the model's performance.

Finally, further improvements could be made by incorporating additional features into the model training. Focusing on addressing the most problematic prediction errors could yield more beneficial improvements.

4. Conclusions

The application of Lithological Layer Prediction (LLP) using machine learning algorithms has significantly improved database validation in mineral exploration in the Paragominas Bauxite Province (PBP). By analyzing and interpreting a vast amount of geological data with machine learning techniques, the accuracy and speed of lithology prediction has been significantly improved. This technology has proven to be a valuable tool in identifying and characterizing bauxite deposits, allowing for more efficient and targeted exploration efforts.

The incorporation of ML in mineral exploration has revolutionize the field, by generating a more comprehensive understanding of the mineral potential in bauxite regions. Future developments in LLP hold immense promise for enhancing geological knowledge and optimizing mineral exploration strategies.

This research serves as a foundation for further advancements in the application of ML and data-driven approaches in the geological field, particularly in the context of lithology prediction and database validation for mineral exploration.

5. References

1. Dramsch, J. S. (2020). 70 years of machine learning in geoscience in review. *Advances in geophysics*, *61*, 1-55.
2. Vladimir Puzyrev, Mario Zelic, Paul Duuring. Applying neural networks-based modelling to the prediction of mineralization: A case-study using the Western Australian Geochemistry (WACHEM) database, Ore Geology Reviews, Volume 152, 2023.
3. Kirkwood, 2016 - A machine learning approach to geochemical mapping https://doi.org/10.1016/j.gexplo.2016.05.003
4. Mohamed, 2019 - Formation Lithology Classification: Insights into Machine Learning Methods https://doi.org/10.2118/196096-MS
5. Kuhn, 2018 - Lithologic mapping using Random Forests applied to geophysical and remote-sensing data: A demonstration study from the Eastern Goldfields of Australia https://doi.org/10.1190/geo2017-0590.1
6. Hearst, 1998 - Support vector machines https://doi.org/10.1109/5254.708428
7. Otchere, 2021 - Application of supervised machine learning paradigms in the prediction of petroleum reservoir properties: Comparative analysis of ANN and SVM models https://doi.org/10.1016/j.petrol.2020.108182
8. Sebtosheikh, 2015 - Lithology prediction by support vector classifiers using inverted seismic attributes data and petrophysical logs as a new approach and investigation of training data set size effect on its performance in a heterogeneous carbonate reservoir https://doi.org/10.1016/j.petrol.2015.08.001
9. Dev, 2019 - Gradient Boosted Decision Trees for Lithology Classification https://doi.org/10.1016/B978-0-12-818597-1.50019-9
10. Martin, 2021 - Centimeter-Scale Lithology and Facies Prediction in Cored Wells Using Machine Learning https://doi.org/10.3389/feart.2021.659611
11. O'Shea, 2015 - An Introduction to Convolutional Neural Networks https://doi.org/10.48550/arXiv.1511.08458
12. Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. MIT Press.
13. LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436–444.
14. Liu, W.B., Wang, Z.D., Liu, X.H., Zengb, N.Y., Liu, Y.R., Alsaadi, F.E., 2017. A survey of deep neural network architectures and their applications. Neurocomputing 234, 11–26.
15. Basile Kotschoubey et al., Caracterização e Gênese dos depósitos de bauxita da Província Bauxitífera de Paragominas, Noroeste da Bacia de Grajaú, Nordeste do Pará/Oeste do Maranhão. In: Marini Onildo João et al. (orgs.). Caracterização de depósitos minerais em distritos mineiros da Amazônia. Brasília, DF, DNPM-CT Mineral, ADIMB, 2005, 613-698.
16. Boslaugh, Sarah and Paul Andrew Watters. 2008. Statistics in a Nutshell: A Desktop Quick Reference, ch. 7. Sebastopol, CA: O'Reilly Media.
17. Pedregosa, 2011 - Scikit-learn: Machine Learning in Python https://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html
18. Chen, 2016 - XGBoost: A Scalable Tree Boosting System https://doi.org/10.48550/arXiv.1603.02754
19. Breiman, L. Bagging predictors. Machine Learning 24, 123–140 (1996). https://doi.org/10.1007/BF00058655
20. He, 2015 - Deep Residual Learning for Image Recognition https://doi.org/10.48550/arXiv.1512.03385
21. Hochreiter, 1997 - Long Short-Term Memory https://doi.org/10.1162/neco.1997.9.8.1735
22. Dwarampudi, M. & Reddy, N. V. S. Effects of padding on LSTMs and CNNs (2019). https://arxiv.org/pdf/1903.07288.
23. Zaharia, 2018 - Accelerating the Machine Learning Lifecycle with MLflow https://cs.stanford.edu/~matei/papers/2018/ieee_mlflow.pdf