

# A Machine Learning Based Approach to Automate Stratigraphic Correlation through Marker Determination

Monthep Parimontonsakul, Somwipa Lotongkum, and Khunalai Mularlee, Chevron Thailand Exploration and Production, Ltd., Bangkok, Thailand

## Summary

Stratigraphic correlation is well recognized as one of the essential processes, providing information regarding stratigraphic and compartmentalization in a reservoir. It becomes a starting point for subsurface evaluation processes ranging from reservoir characteristics to reserves and resources estimation and economic evaluation. It has always been a focus area in numerous traditional and modern research. Several practices approach stratigraphic correlation, including direct tracing from outcrop, relating geological markers, and comparing the organism characteristics. This work focuses only on one of the traditional work processes, utilizing geological markers to identify stratigraphic correlation. This work primarily studies the potential adoption of data analytics and machine learning in identifying geological markers and connecting them to derive stratigraphic correlation.

Well logging information is the primary data source to interpret geological markers. Determining markers was previously done based on the specific well-log characteristics that are rare and uniquely identified in the geological area. It usually takes tremendous effort to find a particular marker from well logging information, especially when many wells scale up the works. Deriving computer-assisted technology with machine learning becomes a key enabler in accelerating and enhancing the business process. The machine learning assisted system has been trained with the entire geoscientists' marker interpretations.

The system consists of two connected machine-learning models. The first model, designed as a multi-class classification, identifies the geological markers using well-logging information. The first model's predicted markers are then fed as an input to the second model, designed as a binary classification. It analyzes the relationship between markers in the same wellbore. Subsequently, the predicted markers resulting from two connected models are linked between two or more wells in the same region to create the stratigraphic correlation. Aiming to determine the practicality and potential adoption from one to another, this study implements the same model concept with two different sets of data, two fields in the Gulf of Thailand. The system has been proven successful in model development and deployment and has achieved nearly human performance levels.

## Introduction

In recent decades, the new asset class, unconventional resources, has been developed to fulfill an increasing need in energy demand. Several activities have been executed in the energy industry, including exploring new oil and gas areas, delineating the additional reservoirs in an existing field, and developing new wells to optimize field production. These activities lead to a significant increase in the amount of data and effort to complete all tasks in the limited time. Stratigraphic correlation is one of the essential processes ranging from an exploration phase to asset development. It provides information regarding stratigraphic and compartmentalization in a reservoir (Howell 1983; Olea and Davis 1986; Waterman and Raymond 1987; Bakke and Griffiths 1989; Fang et al. 1992; Luthi and Bryant 1997).

---

Copyright © the author(s). This work is licensed under a Creative Commons Attribution 4.0 International License.

DOI: 10.14800/IOGR.1204

Received June 2, 2022; revised September 14, 2022; accepted November 14, 2022.

\*Corresponding author: [monthepp@chevron.com](mailto:monthepp@chevron.com)

Stratigraphic correlation is one of the focus areas in numerous traditional and modern research. Rudman and Lankston (1973), Mann and Dowell Jr (1978) identified a stratigraphic correlation, also called correlation, using the cross-correlation technique. Smith and Waterman (1980), Anderson and Gaby (1983), Howell (1983), Waterman and Raymond (1987), Fang et al. (1992), Edwards et al. (2018), Behdad (2019), and Le et al. (2019) determined a similarity between two well-log sequences using the dynamic time warping, also called dynamic waveform matching technique. Zimmermann et al. (2018), Brazell et al. (2019), Bakdi et al. (2020), Tokpanov et al. (2020), and Parimontonsakul (2021) focused on applying machine learning models in stratigraphic correlation identification. This work aims to address the correlation tasks through the geoscientists' and data analytics' lens, synchronizing with the business workflow related to the stratigraphic correlation.

This work presents the use of data analytics and machine learning models to assist or automate the stratigraphic correlation tasks. It focuses only on the traditional work process using a geological marker and explores the potential adoption of data analytic and machine learning models to enhance the process. The same model concept is applied to two fields in the Gulf of Thailand to determine the practicality and potential adoption from one to another.

As a result, the models nearly achieve human performance, supporting the idea of integrating the data analytic workflow into the business workflow. The model results also emphasize the importance of a thorough understanding of the work process through the successful implementation of two models connected in series to improve the model performance. In addition, the two connected models may not easily achieve without the ability to adjust or tweak the model setup, emphasizing the significance of data analytic understanding.

## **Stratigraphic Correlation**

Several kinds of information are required to identify stratigraphic correlation. For example, a similarity in the fossil content can be interpreted as correlative since it presents the same organism characteristics, which infer the same age of the rock units. A similarity in the unique lithology sequences can also be interpreted as correlative since it derives a distinct lithology sequence in the geological area. Given the above examples, this study elaborates that geologists can derive a stratigraphic correlation through geological information: a similarity or specific characteristics between that information, such as lithology, organisms, and a geological period.

Well logging is one of the primary data acquisition processes, providing lithology information, petroleum reservoirs, and petrophysical properties. This information is sufficient to imply a stratigraphic correlation. One of the traditional interpretations that geoscientists usually start with is a marker. A marker, also called a geological marker or horizon, is defined as a rare and uniquely identified lithology sequence that one can map over a geological area (Neuendorf et al. 2011). Connecting the same marker exposed in several wells can imply a stratigraphic correlation because it provides the connection of the unique lithology sequence in the area. However, the correlation interpretation does not guarantee that those correlative reservoirs always have pressure communication since several unknown factors can contribute to compartmentalization, such as fault and unconformity (Parimontonsakul 2021).

As geologists can use well-logging information to interpret a marker, this study proposes that the same process can be assisted by a data analytic process such as machine learning or any computational process. If geologists' marker interpretation is available, the classification model can be implemented using well-logging information as independent variables or features and labeled marker information as the target predicted values. A clustering model can execute when the only available information is the well-logging data, implying that the machine learning model will provide the group of well-logging patterns, aka pseudo markers (Parimontonsakul, 2021). This study focuses only on the classification model where geologists' marker interpretation is available in this work. This study applies the work process to two fields in the Gulf of Thailand to determine the use case and application in the actual field data. Two areas are selected as it provides the evidence to demonstrate the applicability of the workflow to other fields.

## Analytical Problem Formulation

As previously mentioned, marker identification is an essential process in stratigraphic correlation. This section elaborates on the analytical problem formulation and the model implication. This study intends to focus on only one analytical formulation, Convolutional Neural Network, to demonstrate the implementation of the work process in the actual field data.

Convolutional Neural Network (CNN) model has become predominant in the image recognition and computer vision research areas, including several architectures aiming to achieve higher accuracy and more efficient calculations (Krizhevsky et al. 2012; Krizhevsky et al. 2017; Wang et al. 2020; Parimontonsakul 2021). Due to the limited computational capability, this study employs one of the simplest and the most efficient architectures, MobileNets, to identify the markers. Parimontonsakul (2021) proposes one transformation of well-logging information to the image approach. The primary concept is that one column vector represents one section of a well-log sequence, while well-log interval and compression factors, implying zoom-in and out of an image, are added to the well-log sequence. This process creates additional column vectors to the same well log series. He initially proposes to apply eight factors to the same well log sequence, resulting in eight vectors of an image, defined as a pad. The same process is used in other well-log sequences to create the complete image.

Applying the same concept, this study can make several well-logging images by changing the well-log interval and compression factors, as illustrated in **Figure 1**. The well-log series from left to right are the gamma-ray log, resistivity log, neutron log, and density log in the sandstone scale. The transformation of a well-log to an image is demonstrated. **Figure 2** presents the MobileNets architecture. The transformed well log image is set up as an input layer in the CNN model, while the output layer is the geoscientists' marker interpretations. Since there is more than one marker in an output layer, the problem is set up as a multi-class classification problem.

As the reader is aware that the transformed image consists of only a well log sequence, there is no additional information regarding well location, well depth, or any other information useful in identifying the geological marker. This study introduces two approaches to resolve the concern. The first approach is to create a placeholder column vector in the image as a place to input supplemental information such as well location and well depth. This method provides the simplest solution since it does not require any adjustment to the MobileNets architecture. The second approach is to include additional data input connecting directly to the fully connected layer. In this case, the modified MobileNets architecture needs to be constructed to add extra information to the CNN model. The two approaches are illustrated in **Figure 3**. In this work, the first approach is applied for simplicity in model creation, but this does not imply that the first approach is the best. As a result, this study proposes the modified MobileNets architecture as demonstrated in **Table 1**.

This study addresses this issue by creating an additional machine learning model that explores the probability of correctness in the marker prediction given the nearby marker prediction information. This process can be viewed simply as the nearby marker prediction should be similar, or if the marker is alphabetically sorted from shallow to deep, the nearby marker prediction will also be alphabetically sorted in the same manner. There are several techniques to formulate this kind of model. One of the approaches is to formulate it as a classification problem, where independent variables are the predicted markers from the CNN model, including predicted marker depth and other mathematical aggregation such as min, mean, mode, and a max of the numerical values, and the correctness of the marker prediction given the depth threshold as the target predicted values.

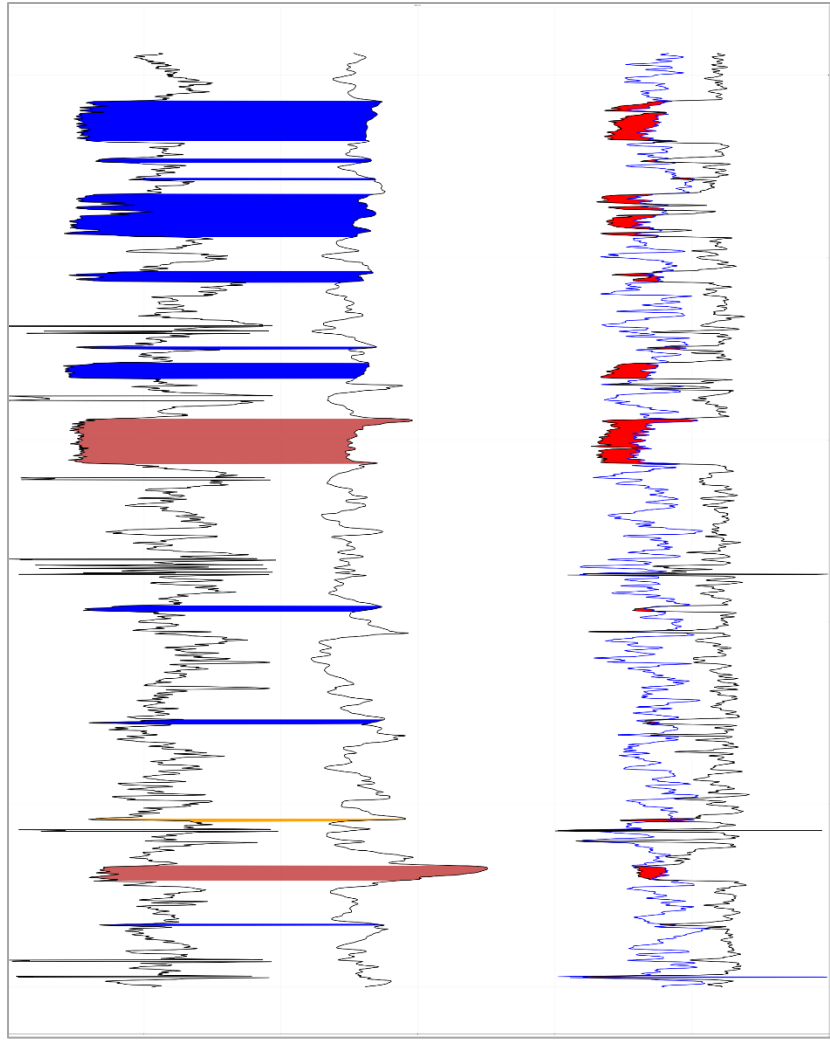


Figure 1—The well log images for MobileNets (Parimontonsakul 2021).

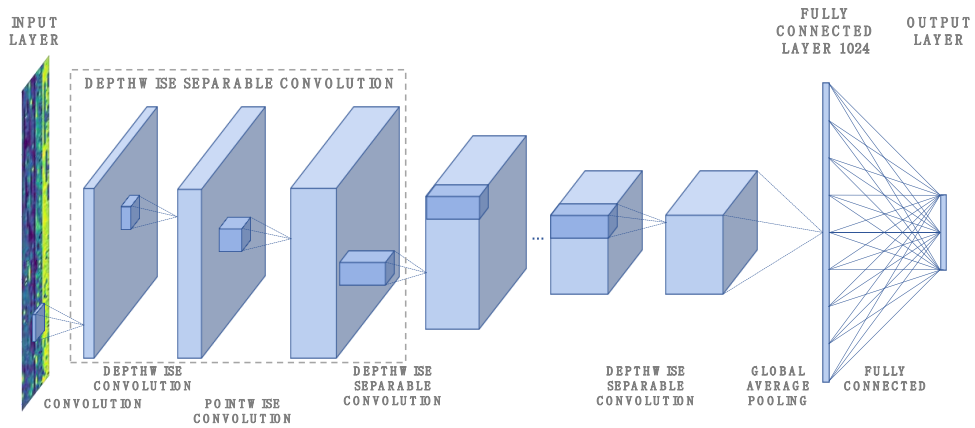
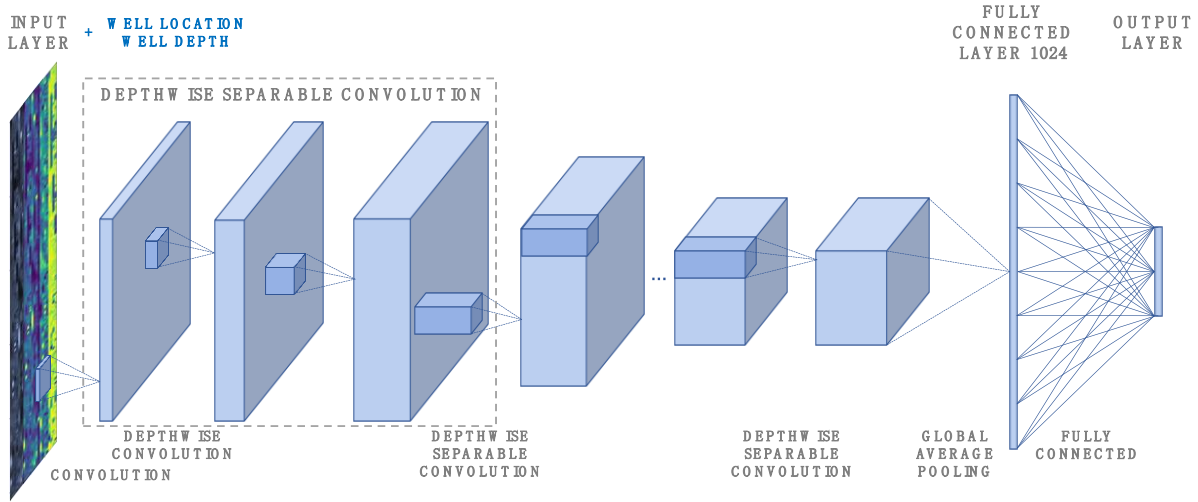
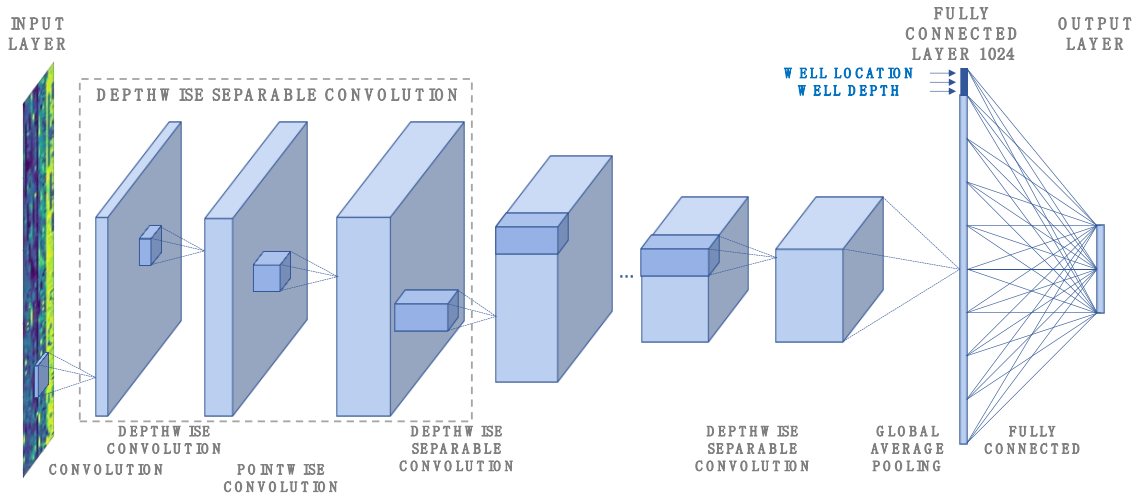


Figure 2—The MobileNets architecture (Parimontonsakul 2021).



(a)



(b)

**Figure 3—The modified MobileNets architecture alternative for additional data input (modified from Parimontonsakul 2021). (a) The modified MobileNets architecture added additional column vector; (b) The modified MobileNets architecture added additional data input to fully connected layer.**

## Metrics

This section discusses the metrics to validate the model performance. This study presents metrics for the 3-class classification problem, as illustrated in **Table 2**; however, the same formula can apply to the multi-class classification problem. The *precision* is the number of correctly classified divided by all predicted values with the same condition, while the *recall* is the number of correctly classified divided by all true values with the same condition. *F1-score* is another metric usually applied in classification problems. It can be written as a harmonic means of precision and recall. The mathematical expressions of precision, recall, and F1-score are shown in **Eqs. 1 to 3** (Sokolova and Lapalme 2009). In this work, the model prediction performance in each marker is measured by F1-score, while the macro average of the F1-score measures the overall model prediction performance.

$$Precision_i = \frac{tp_i}{tp_i + \sum_j fp_{ij}}, \dots \dots \dots (1)$$

$$Recall_i = \frac{tp_i}{tp_i + \sum_j fn_{ij}}, \dots \dots \dots (2)$$

$$F1-score_i = \frac{2 \times Precision_i \times Recall_i}{Precision_i + Recall_i} \dots \dots \dots (3)$$

**Table 1—The proposed modified MobileNets architecture.**

Layer	Type	Output Shape
input	Input Layer	(512, 36, 1)
scal_input	Scaled Input Layer	(256, 32, 1)
conv1	Conv2D	(128, 16, 32)
conv1_bn	BatchNormalization	(128, 16, 32)
conv1_relu	ReLU	(128, 16, 32)
conv_dw_1	DepthwiseConv2D – BatchNormalization - ReLU	(128, 16, 32)
conv_pw_1	Conv2D – BatchNormalization - ReLU	(128, 16, 64)
conv_dw_2	DepthwiseConv2D – BatchNormalization - ReLU	(64, 8, 64)
conv_pw_2	Conv2D – BatchNormalization - ReLU	(64, 8, 128)
...	...	...
conv_dw_13	DepthwiseConv2D – BatchNormalization - ReLU	(8, 1, 1024)
conv_pw_13	Conv2D – BatchNormalization - ReLU	(8, 1, 1024)
avg_pool	GlobalAveragePooling2D	(1024)
fc	FullyConnected	(1024, 14)
output	Output Layer	(14)

**Table 2—The 3-class classification problem confusion matrix.**

		True Condition		
		Condition A	Condition B	Condition C
Predicted Condition	Condition A	$tp_a$	$fp_{ab} = fn_{ba}$	$fp_{ac} = fn_{ca}$
	Condition B	$fp_{ba} = fn_{ab}$	$tp_b$	$fp_{bc} = fn_{cb}$
	Condition C	$fp_{ca} = fn_{ac}$	$fp_{cb} = fn_{bc}$	$tp_c$

## Model Development and Deployment

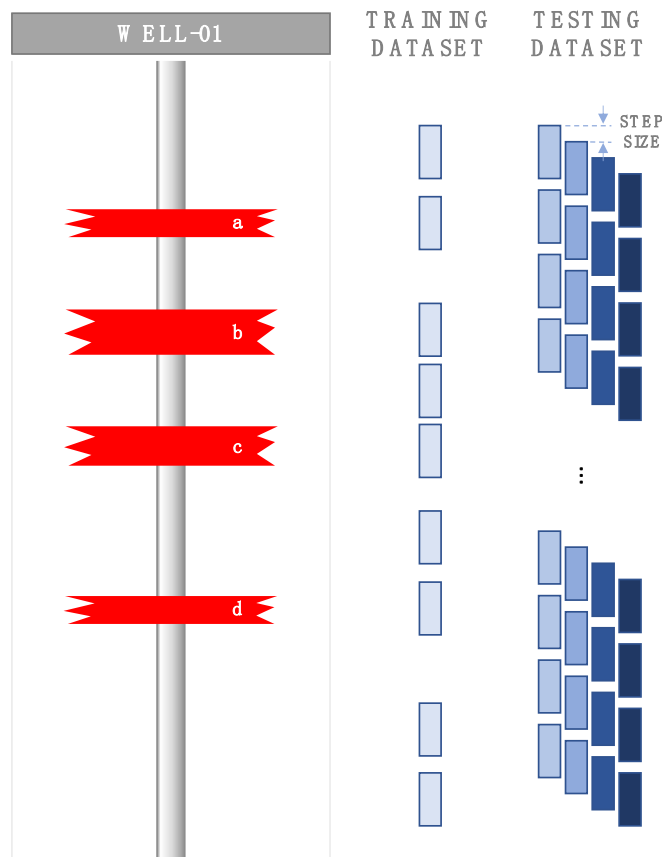
Figures 4 and 5 demonstrate two processes in model development and deployment for each model. The first process focuses on the first model, identifying the geological markers, development and deployment, while the second process focuses on the second model, analyzing the relationship between markers in the same wellbore, development, and deployment.

During the first model development, called as marker prediction model, the marker and non-marker data are required as input to the model. The imbalanced dataset from significantly different amounts of data between marker and non-marker needs to be addressed. This work applies the simple sampling approach to reduce the non-marker data to reduce the imbalanced dataset and control the computational time and memory to stay within the acceptable range.

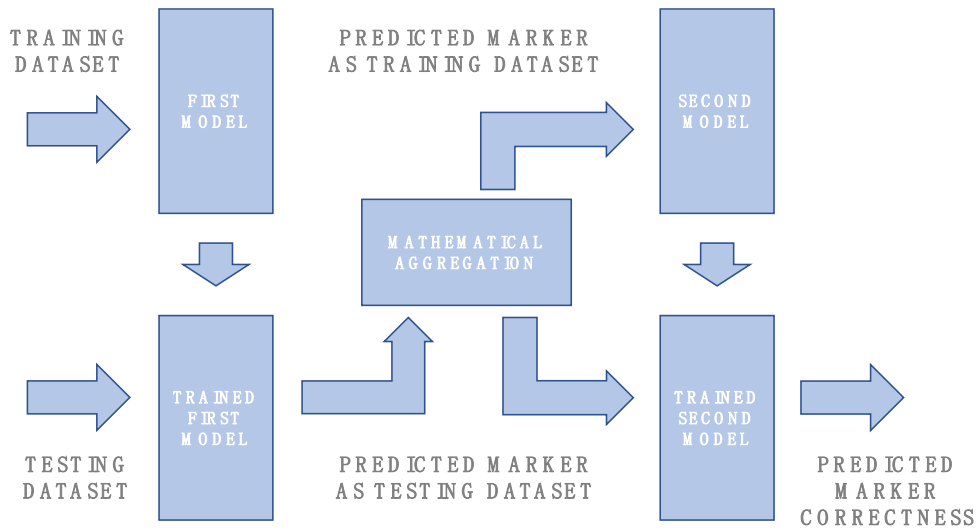
During the marker prediction model deployment, each interval of the well-log sequence is evaluated with the trained machine learning model defined in the previous step. The step size between each evaluation is a major concern. If the step size is too small, the number of function evaluations will be high, leading to higher computational time and vice versa. In addition, if the step size is too big, it also negatively impacts the model performance of the second process, implying that the second process will be less accurate. This study does not recommend any concrete solutions that always demonstrate the best step size and suggest further study in this area as needed.

In the second model, named the marker association model, development and deployment process are straightforward as there is no other process before the model deployment, as demonstrated in the marker prediction model.

Two fields in the Gulf of Thailand, classified as fluvial depositional environments, are evaluated (**Table 3**). Geoscientists can identify the geological markers in both fields as the initial set of data to perform the correlation tasks. However, the markers identified by each area are not the same. Since each location may have a different set of marker interpretations and originations, this study intends to develop and deploy the model in its area to honor its data distribution. Field A and Field B are trained and validated separately with the same model concept to demonstrate that the same process can apply in other locations without any adjustment. Less than half of the field data have been trained due to limited computational performance, while the models have been tested in all remaining data.



**Figure 4—The training and testing dataset definition.**



**Figure 5—The two connected models architecture.**

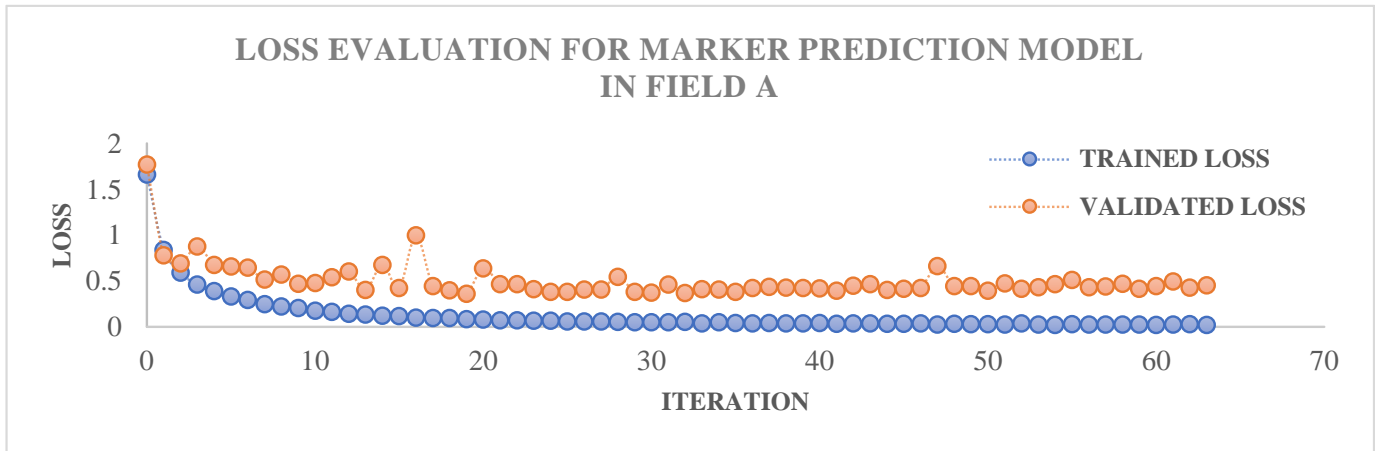
**Table 3—The validated metrics of the MobileNets model in two field in the Gulf of Thailand.**

Marker	Field A			Field B		
	Precision	Recall	F1-score	Precision	Recall	F1-score
C	0.9240	0.9356	0.9298	0.9412	0.9287	0.9349
D	0.9392	0.9727	0.9557	0.9276	0.9899	0.9578
E	0.9118	0.9602	0.9354	0.9225	0.9551	0.9385
F	0.9544	0.9458	0.9501	0.9364	0.9346	0.9355
G	0.9298	0.9620	0.9456	0.9233	0.9736	0.9478
H	0.9484	0.9682	0.9582	0.9365	0.9669	0.9514
I	0.9632	0.9458	0.9545	0.9810	0.9513	0.9659
J	0.9332	0.9508	0.9419	0.9159	0.9656	0.9401
K	0.9159	0.9619	0.9384	0.9099	0.9647	0.9365
L	0.9212	0.9509	0.9358	0.9037	0.9494	0.9260
M	0.8370	0.9743	0.9004	0.8477	0.9905	0.9136
N	0.9013	0.9211	0.9111	0.8826	0.9259	0.9037
O	0.9600	0.9796	0.9697	0.9506	0.9647	0.9576
NA	0.8162	0.7218	0.7661	0.7963	0.7176	0.7549
macro avg	0.9183	0.9393	0.9280	0.9125	0.9413	0.9267



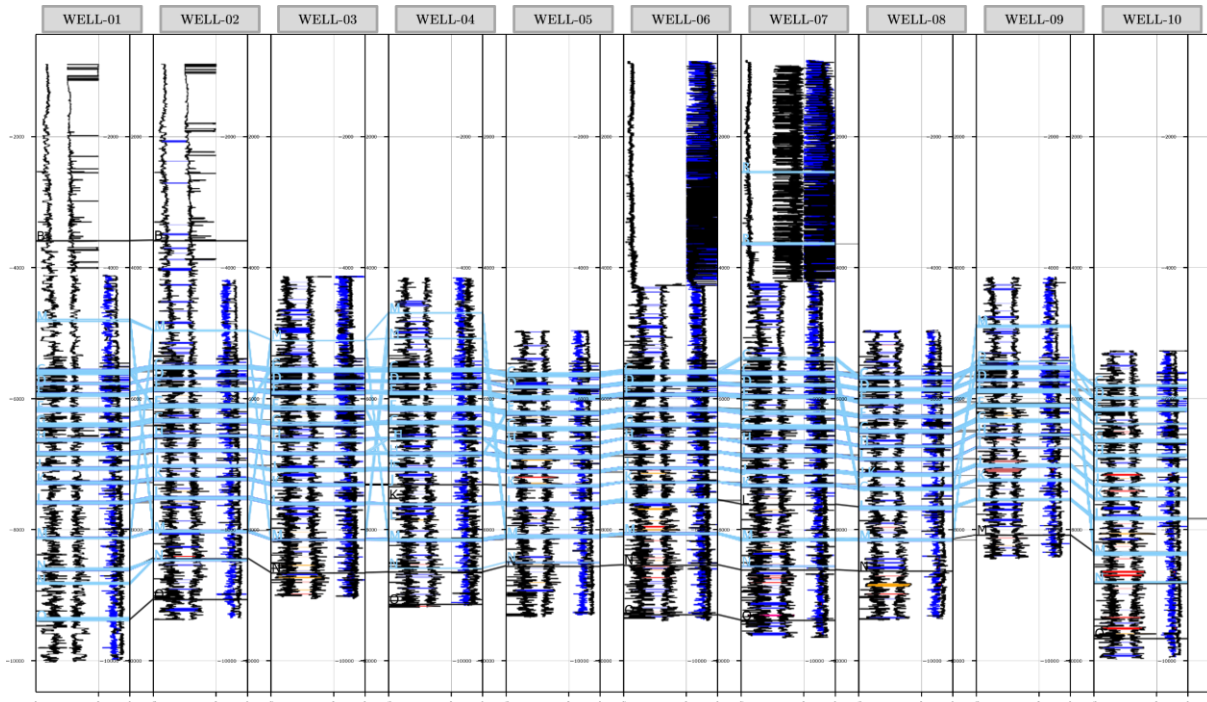
## Discussion

The marker prediction models, the CNN models, yield an incredible performance. The trained model in fields A and B in the Gulf of Thailand presents comparable performance. The F1-score metric reaches up to 0.95 in most markers except the markers *L*, *M*, *N*, *O*, and *NA* in field A and field B. This result demonstrates that the same model concept can be applied in other fields without further adjustment, emphasizing that only the training data of the target field is required to implement the same model concept in other areas. The trained and validated loss function evaluation is demonstrated in **Figure 6**.

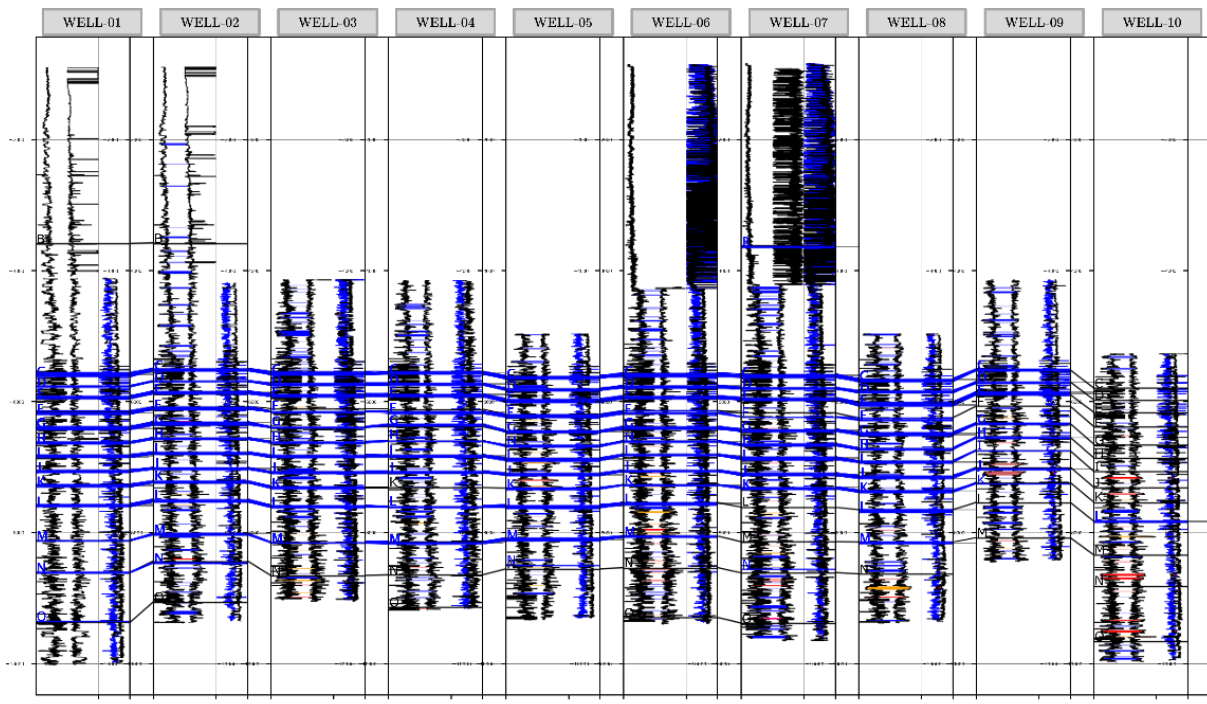


**Figure 6—The loss function evaluation in trained and validated data set in field A.**

**Figure 7** demonstrates that applying the marker prediction model alone in the deployment phase confirms the correlation chaos without the marker association model, as presented as the solid light blue line. Integrating the marker association model deems necessary as it significantly reduces the number of incorrect correlation identifications, as demonstrated by the solid blue line. It also indicates that the marker association model is important in the process as long as the marker prediction model's precision scores do not reach the perfect score. The marker association model will complement and help reduce the marker prediction model's error. It also hypothesizes further optimization in the marker prediction model to provide a higher recall score so that the model tended to interpret more markers and applied the marker association model to remove the incorrectly predicted markers. However, this optimization idea has not been tested yet.



(a)



(b)

Figure 7—The stratigraphic correlation results from the models, the solid light blue line represents the results from the marker prediction model, The solid blue line represents the results from the marker association model. (a)The stratigraphic correlation results from the marker prediction model, elaborating few incorrect correlation identifications from connecting identical markers;(b)The stratigraphic correlation results from the marker prediction and marker association model, presenting an improved correlation identifications by complementing the error from marker prediction model.

## Conclusions

This work introduces a significant step toward the automated stratigraphic correlation using machine learning models in the marker horizon determination problem. The CNN model has been proven successful in identifying the marker horizon in two fields in the Gulf of Thailand. The model performs as it designs and nearly achieves human performance. However, the CNN model alone does not perform as well as it establishes during the model deployment phase. This concern has been proven and the additional machine learning models are connected in series to improve the model performance. The model results highlight a significant improvement in automated stratigraphic correlations.

Additionally, the model concept is proven to deploy in two fields in the Gulf of Thailand without further adjustment, emphasizing the potential adoption of the model concept from one area to another. Given the model concept adoption, the rapid growth in digital transformation applying machine learning and other data analytics in the business workflow will not be a far future. This study concludes that data analytics can integrate with the business process to improve accuracy in stratigraphic correlation and assist geoscientists throughout the process. A thorough understanding of stratigraphic correlation and data analytics empowered by machine learning models is non-trivial and demonstrates a step toward a fully automated system.

## Nomenclature

CNN	=	Convolutional Neural Network
FN	=	False Negative
FP	=	False Positive
TP	=	True Positive

## Acknowledgments

The authors would like to acknowledge Saran Boonchirdchoo and Juntra Wanwises for their efforts in data gathering and explaining the data structure. The authors also would like to express pleasure to Nualjun Kitvarayut for continuous support in the research topic.

## Conflicting Interests

The author(s) declare that they have no Conflicting interests.

## References

- Anderson, K.R. and Gaby, J.E. 1983. Dynamic waveform matching. *Information Sciences* **31**(3), 221-242.
- Bakdi, S., Kannan, N., Reddy Masini, S., et al. 2020. Automated Well Correlation using Machine Learning and Facial Recognition Techniques. Paper presented at the Abu Dhabi International Petroleum Exhibition & Conference, 9-12 November. SPE-203301-MS.
- Bakke, S. and Griffiths, C.M. 1989. Interactive Quantitative Matching of Stratigraphic Sequences of Numerical Lithostates Based on Gene-Typing Techniques. In *Correlation in Hydrocarbon Exploration*, ed. Collinson, J.D, Chap. 5, 61-76. Dordrecht: Springer.
- Behdad, A. 2019. A Step Toward the Practical Stratigraphic Automatic Correlation of Well Logs Using Continuous Wavelet Transform and Dynamic Time Warping Technique. *Journal of Applied Geophysics* **167**: 26-32.
- Brazell, S., Bayeh, A., Ashby, M., et al. 2019. A Machine-Learning-Based Approach to Assistive Well-Log Correlation. *Petrophysics* **60**(4): 469-479.
- Edwards, J., Lallier, F., Caumon, G., et al. 2018. Uncertainty Management in Stratigraphic Well Correlation and Stratigraphic Architectures: A Training-Based Method. *Computers and Geosciences* **111**:1-17.

- Fang, J., Chen, H., Shultz, A.W., et al. 1992. Computer-Aided Well Log Correlation. *AAPG Bulletin* **76**(3): 307-317.
- Howell, J. 1983. A FORTRAN 77 Program for Automatic Stratigraphic Correlation. *Computers and Geosciences* **9**(3): 311-327.
- Krizhevsky, A., Sutskever, I., and Hinton, G.E. 2012. Imagenet Classification with Deep Convolutional Neural Networks. *Advances In Neural Information Processing Systems* **25**:1097-1105.
- Krizhevsky, A., Sutskever, I., and Hinton, G.E. 2017. ImageNet Classification with Deep Convolutional Neural Networks. *Communications of the ACM* **60**(6):84-90.
- Le, T., Liang, L., Zimmermann, T., et al. 2019. A Machine Learning Framework for Automating Well Log Depth Matching. *Petrophysics* **60**(5):585-595.
- Luthi, S.M. and Bryant, I.D. 1997. Well-log Correlation Using a Back-Propagation Neural Network. *Mathematical Geology* **29**(3):413-425.
- Mann, C. J. and Dowell Jr, T.P. 1978. Quantitative Lithostratigraphic Correlation of Subsurface Sequences. *Computers and Geosciences* **4**(3):295-306.
- Neuendorf, K.K., Mehl Jr., J.P., and Jackson, J.A. 2011. *Glossary of Geology (5th Edition)*. American Geosciences Institute.
- Olea, R.A. and Davis, J.C. 1986. An Artificial Intelligence Approach to Lithostratigraphic Correlation Using Geophysical Well Logs. Paper presented at the SPE Annual Technical Conference and Exhibition, 5-8 October. SPE-15603-MS.
- Parimontonsakul, M. 2021. An Analytical Approach to Automate Stratigraphic Correlation using Well Logging Information. PhD dissertation, Massachusetts Institute of Technology, Cambridge, Massachusetts.
- Rudman, A. J. and Lankston, R.W. 1973. Stratigraphic Correlation of Well Logs by Computer Techniques. *AAPG Bulletin* **57**(3): 577-588.
- Smith, T. F. and Waterman, M.S. 1980. New Stratigraphic Correlation Techniques. *The Journal of Geology* **88**(4): 451-457.
- Sokolova, M. and Lapalme, G. 2009. A Systematic Analysis of Performance Measures for Classification Tasks. *Information Processing & Management* **45**(4): 427-437.
- Tokpanov, Y., Smith, J., Ma, Z., et al. 2020. Deep-Learning-Based Automated Stratigraphic Correlation. Paper presented at the SPE Annual Technical Conference and Exhibition, 26-29 October. SPE-201459-MS.
- Wang, W., Li, Y., Zou, T., et al. 2020. A Novel Image Classification Approach via Dense-MobileNet Models. *Mobile Information Systems* **20**:1-8
- Waterman, M.S. and Raymond, R. 1987. The Match Game: New Stratigraphic Correlation Algorithms. *Mathematical Geology* **19**(2): 109-127.
- Zimmermann, T., Liang, L., and Zeroug, S. 2018. Machine-Learning-Based Automatic Well-Log Depth Matching. *Petrophysics* **59**(6): 863-872.