



**Lee Ming Xiang** ✉ [mingxiang1006@gmail.com](mailto:mingxiang1006@gmail.com)

☎ (+60)12-9791319

I am a domain data scientist with petroleum geoscience background. I have six years of comprehensive experience in developing and deploying AI solutions to address challenges within oil and gas industry.

Proficient in natural language processing (NLP), GenAI LLM, subsurface relevant AI solutions. I have successfully implemented AI-driven strategies to enhance operational efficiency and decision making. [LinkedIn profile](#)



## Core Skills

Machine Learning,  
Deep Learning, Computer  
Vision, Natural Language  
Processing, GenAI LLM,  
Petroleum Geoscience

## Data Science Skills

Python, Pytorch, Pytorch  
Lighting, TensorFlow, Scikit-  
Learn, MLflow,  
Spotfire, Power BI, Dash,  
Azure, Structure Query  
Language (SQL),  
Oracle Database,  
Docker, Dataiku

## Geoscience Skills

Petrel, Omega, Vista, Techlog,  
Rock Physics, Seismic  
Processing, Seismic  
Interpretation, Static  
Modelling, Exploration &  
Production Cycle

## Research Skills

Problem Solving,  
Critical Thinking, Innovative,  
Collaborative, Rapid  
Assimilation

## Working Experience

### Senior Domain Data Scientist


*KL Innovation Factory, SLB*

*Jun 2021 – Present*


*Kuala Lumpur, Malaysia*


- 1. ResNet Driven On-Prem OCR for Handwritten Document Analysis**
  - Engineered **Optical Character Recognition (OCR) segmentation** techniques, including line level and word level segmentation, to enhance text detection accuracy and efficiency.
  - Finetuning ResNet-72 architecture**, combining **Long Short-Term Memory (LSTM)** layers in the network, and **achieved Character Error Rate (CER) ~17%**.
- 2. High-Fidelity Pre-stack Seismic Signal Enhancement with Self-Innovate Network Architecture**
  - Low-Frequency Self-Supervised Sequential Inference Model (LoF-SSIM)**: developed innovative self-supervised learning model, with sequential inferencing technique for swell noise attenuation.
  - Low-Frequency Hybrid Resnet-ECA Network**: designed and implemented a hybrid network combining ResNet with Efficient Channel Attention (ECA) for effective linear noise removal.
  - Achieved superior performance compared to traditional seismic processing methods**, delivering highly consistent, noise-free results with minimal signal leakage and artifacts.
- 3. Cross-Domain Operational Intelligence Retrieval with Advanced LLM Framework [\[Published soon\]](#)**
  - Developed **comprehensive LLM framework** integrating domain prompt engineering, Retrieval Augmented Generation (RAG), and fine-tuning of large language models (LLMs).
  - Enhanced model performance** by fine-tuning an open-source LLM, achieving a 94% accuracy rate—**surpassing finetuned GPT-3.5 Turbo by 3%, pretrained One-Shot GPT4 by 13%**—through strategic hyperparameter tuning and rigorous post-processing safeguards.
  - Conducted **extensive research** on chunking strategies, data types, embedding techniques, and vector database within the RAG framework, leading to a refined and highly effective domain insights retrieval system. This research **addressed technical abbreviations and challenges associated with outdated domain knowledge**.
  - Versatile LLM framework for multi-domain application**, including drilling engineer for offset well analysis, geomechanics engineer for post-drill model assessment, and reservoir engineer for strategic well placement planning, **demonstrating impact in diverse operational contexts**.
- 4. Real-Time Distributed for Fiber Optic Sensing with Edge-Enabled Streaming Analysis**
  - Revolutionized Distributed Acoustic Sensing (DAS) data processing by **developing a novel signal-to-image conversion algorithm**. Transforming >600,000 data points of high-frequency DAS signals into compact image representations, **achieving a 71% reduction (14MB to 4MB)** in data size while preserving critical signal characteristics. Estimated **bandwidth cost saving ~10,000 USD per month**.
  - Engineered cutting-edge anomaly detection system using **Self-supervised Transformer with Energy-based Graph Optimization (STEGO)**. Adapted the STEGO architecture for time-series data, incorporating attention mechanisms and self-supervised pretraining on domain-specific unlabeled data, delivering a robust solution of **unsupervised semantic segmentation in anomaly detection**.
- 5. Probabilistic Rock Physics Modelling via ML Optimization [\[Published soon\]](#)**
  - Integrated **global sensitivity analysis, derivative free local optimization, and ML clustering**, with theoretical rock physics bounds constraint, to generate stochastic and actionable models
  - Deployed human centric AI SaaS application** integrating Azure function (FaaS) to retrieve data from Postgres SQL (self-designed and deployed), SendGrid API for email services to engineer.
- 6. Transfer-Learned Language Model for Upstream Petroleum Application [\[Link\]](#) [\[Link\]](#) [\[Link\]](#)**
  - Developed and **fine-tuned pre-trained language model tailored for upstream petroleum applications**, to generate an **unsupervised multitask learning model** for downstream NLP tasks.
  - Demonstrate the value fine tuning domain specific language model **surpassing pretrained GPT-2 small and medium model by 10% of accuracy**.

## Certifications

 **Generative AI with Large Language Models**

 **Build Basic GANs**

 **Improving Deep Neural Networks: Hyperparameter Tuning**

 **Neural Networks and Deep Learning**

 **Dataiku Core Designer**

 **Dataiku ML Practitioner**

 **Dataiku Advanced Designer**

 **Tibco Certified Associate Spotfire**

 **Industrial Data Fundamentals**

 **Data Fusion Fundamentals**

 **OSDU Developer Training**

 **Oracle Database Design & Programming with SQL**

 **Azure AI & Data Fundamentals**

 **Geosolutions Horizon Fixed Step Training Phase 1 ,2,3**

## 7. Cross Domain Behind Casing Opportunities (BCO) – Petrophysics & Production

- Engineered a **robust analytical framework for cross-domain pattern recognition** using **Explainable AI techniques and ensemble LightGBM** models. Utilized SHAP and LIME for model interpretability, and Shapash, Explainer, and Dalex Arena Dashboard for comprehensive domain analysis.
- Integrated multi-modal petrophysical data (e.g., resistivity, porosity, permeability) with time-series production rates for a field with 90 wells and 239 well zones, **achieve F1 score of 74% for perforation flag, 95% for BCO flag, 74% for permeability prediction, and 80% of Net Pay Permeability**.

## 8. Adaptive Prescriptive Well Performance Intelligence Application [[Publication Section Item 4](#)]

- Developed a prescriptive analytics dashboard for well performance, **integrating multi-sources data** (well status, monthly fluid rate, hourly operation data, well test data, monthly fluid target, and well metadata) **with ~ 3 million data points**, for a comprehensive field and well level analysis.
- Optimized performance in big data analytics, using batch splitting, asynchronous I/O **achieving 87% faster data retrieval**, and multiprocessing approach for **63% faster processing**.

## 9. Predictive ML for Real-Time CO2 Emission Forecasting [[Link](#)]

- Developed and deployed a **predictive machine learning system** for CO2 emission forecasting based on gas fuel rate predictions. Implemented **autonomous model selection** to choose the optimal model from a series of tree-based and transformer models for each field.
- Developed a custom data pipeline that **integrated the ML system with Power BI using RESTful API**, enabling near real-time CO2 footprint monitoring across the basin.

## Geophysicist

Geosolutions, SLB

Mar 2018 – May 2021

Kuala Lumpur, Malaysia

Seismic processing geophysicist, with experience in narrow and wide azimuth surveys at ultra-shallow, shallow, and deep water in different offshore basins. Experience in seismic processing, mainly on seismic deblending, seismic denoise, demultiple, migration, and velocity picking.

- Conventional P190 information extraction & SEG-D file list generation were done sequence by sequence has consumed significant production time. Developed an automated bash Linux script for automatic data extraction, **saving production from 2 weeks to 20 minutes**.
- Automated Parameter Analysis and Recommendation for Adaptive Deghosting, **saving testing time about 70%**.

## Publications

### ➤ [Data Science](#)

- Retrieving Operation Insights with GenAI LLM: Comparative Analysis and Workflow Enhancement, M.X. Lee, Z. Wang, ADIPEC, Abu Dhabi, UAE, November 2024.
- Integrating Data Science into Rock Physics Modelling Workflow, M.X. Lee, I. Kozlov, GEO 4.0 Digitalization in Geoscience Symposium, Al Khobar, Saudi Arabia, October 2024.
- Unsupervised Multitask Learning for Oil and Gas Language Models with Limited Resources, M. Marlot., D.N. Srivastava, F.K. Wong, M.X. Lee, ADIPEC, Abu Dhabi, UAE, October 2023. [[Link](#)]
- Optimizing Performance in Big Data Handling for Enhanced Data Analytics, S. Atiq, M.X. Lee, EAGE Workshop on Data Science, 2023.
- A survey of Natural Language Processing in Oil and Gas: Opportunities and Challenges, M. Marlot, M.X. Lee, EAGE Workshop on Data Science, 2023.
- Unlocking Value from Text: Visualizing Insights with Natural Language Processing in Unstructured Oil and Gas Reports, M. Marlot, M.X. Lee, A. Irfan, P.K., Tellapaneni, L. Edwin, SPE/IATMI Asia Pacific Oil & Gas Conference and Exhibition, 2023. [[Link](#)]
- Information Retrieval from Oil and Gas Unstructured Data with Contextualized Framework, M.X. Lee, M. Marlot, Third EAGE Digitalization Conference and Exhibition, Mar 2023. [[Link](#)]
- Carbon Dioxide Emission Monitoring based on Prediction of Gas Fuel Rate using Machine Learning, M.X. Lee, EAGE Conference on Digital Innovation for a Sustainable Future, 2022. [[Link](#)]

### ➤ [Geoscience](#)

- Adaptive Deghosting Dashboard, M.X. Lee, A. Sazykin, SLB Technical Coordinators Meeting, 2021.
- Imaging multi-order multiples - Shallow Water Case Study from Southeast Asia, B. Chowdhury, A. Sazykin, P. Kristiansen, S.Y. Lee, M.X. Lee, R. Alai, M. Shah, M. Nasrul, N. Nadzirah, SEG Kuwait "Seismic Multiples - The Challenges and the Way Forward" Workshop, 2019.
- Abstract of Optimum Notch Frequency Recovery using non-CMS approach. C.M. Lam, A. Verba, M.X. Lee, SLB Technical Coordinators Meeting, 2019.
- Application of Simultaneous Inversion Characterizing Reservoir Properties in X Field, Sabah Basin, M.X. Lee, L.A. Luluan, IOP Conference Series: Earth and Environmental Science, Volume 88, 5th International Conferences on Geological, Geographical, Aerospace and Earth Sciences 2017 (5th AeroEarth 2017) 20–21 May 2017, Kuta, Bali, Indonesia. [[Link](#)]

## **Achievements**

- **2nd runner up** of EAGE Field Challenge 2017  
Represented Malaysia participating the EAGE Field Challenge 2017 organized by Total company at Paris, France with fully Integrated evaluation and field development project.
- **AAPG L. Austin Weeks Recipient 2017**  
Scholarship recipient for the 2017 American Association of Petroleum Geologists Foundation L. Austin Weeks Scholarship program
- **Silver Award in Integrated Exploration and Production Opportunity Evaluation Project 2016**  
New prospect finding for Bundi field integrating G&G knowledge. Performed reservoir probability and risk evaluation, and petroleum economic analysis.

## **Mentoring and Leadership**

**International Petroleum Technology Conference 2025 Committee** *Jun 2023 – Feb 2025*  
Participate in organizing the Digital, Data Analytics, and Automation program, review submitted abstract for technical session.

**Technical Committee for SLB Machine Learning Innovative Competition** *Jan 2023 - Jun 2023*  
Review the data science challenges and DELFI technology stack used for the competition.

**Technical Committee & Mentor for APGCE GeoHackathon** *July 2022 - Nov 2022*  
Worked with Petronas management, geoscientists, and data scientist in developing oil and gas upstream data science challenges. Mentoring participants in applying data science to domain challenges.

**Technical Committee for SLB Beijing Geoscience Center** *Dec 2021 - Jan 2022*  
Introduced the hackathon challenge in forecasting the production decline curve using both production and formation data.

**Volunteer Speaker for Women Who Code Power BI Workshop** *Jun 2019*

## **Personal Projects**

- Unsupervised Segmentation using Computer Vision  
[https://github.com/mingxiang1006/Unsupervised\\_Seg](https://github.com/mingxiang1006/Unsupervised_Seg)
- Automatic Detection of Solar Roof Top using Computer Vision  
[https://github.com/mingxiang1006/solar\\_ai](https://github.com/mingxiang1006/solar_ai)
- Groove Defect Segmentation using Computer Vision  
<https://www.kaggle.com/code/mingxiang1006/unet-seg>
- Machine Learning with Optimized Parameters for Ecommerce Product Classification  
<https://github.com/mingxiang1006/Ecommerce-Product-Classification/tree/main>
- Future Sales Prediction  
[https://github.com/mingxiang1006/Predict\\_Futre\\_Sale](https://github.com/mingxiang1006/Predict_Futre_Sale)
- Telco Customer Churn Prediction  
<https://github.com/mingxiang1006/Telco-Customer-Churn-Prediction>
- Nasdaq Stock Portfolio Optimization  
<https://github.com/mingxiang1006/Stock-Portfolio-Optimization>

## **Education**

### **Master of Data Science**

*Oct 2020 – Jun 2022* *University Malaya (UM), Kuala Lumpur*  
Master Thesis: Generation of Carbon Dioxide Emission based on Prediction of Gas Fuel Rate using Machine Learning (Time Series Prediction)

### **Bachelor of Technology (Hons) in Petroleum Geoscience**

*May 2012 – May 2017* *University Technology PETRONAS (UTP), Perak*  
Majoring in Exploration Geophysics, Fundamental in Geology, Petrophysics, and GIS  
Final Year Project: Application of Simultaneous Inversion in Sarawak Basin, Malaysia

### **Student Exchange Program**

*Aug 2015 – Dec 2015* *Missouri University Science & Technology, United States*  
Studied Petroleum Economics, Reservoir Characterization, General Psychology and Technical Communication