**Sử dụng phương pháp truyền thống,học máy để ước lượng độ rỗng**

Trần Nguyễn Thiện Tâma, b, \*, Đặng Xuân Phúa, b

*a Khoa Kỹ thuật Địa chất và Dầu khí,****Trường Đại học Bách Khoa Thành phố Hồ Chí Minh***

***b Đại học Quốc gia Thành phố Hồ Chí Minh***

**\* Tác giả liên hệ:**Đặng Xuân Phú , **Email:**phu.dangxuan268@hcmut.edu.vn; **Điện thoại:** 0767385934

# Tóm tắt

*Độ rỗng đóng vai trò quan trọng nhất trong việc xác định đặc tính của vỉa để đánh giá tiềm năng của vỉa chứa và tối ưu hóa khai thác dầu khí. Vì vậy, ước lượng chính xác độ rỗng là rất cần thiết. Hiện nay, so với các phương pháp tính toán truyền thống, phương pháp học máy đã chứng tỏ được những ưu điểm vượt trội của mình. Trong nghiên cứu này, nhóm tác giả sử dụng các phương pháp học máy phổ biến như* *Multiple linear Regression, Random Forest Regressor, Support Vector Machine,để tính toán độ rỗng thông qua dữ liệu vật lý vỉa. Các kết quả ước lượng sau đó được so sánh để đánh giá và lựa chọn phương pháp tối ưu.*

**Từ khóa:** Độ rỗng, học máy, Multiple linear Regression, Random Forest Regressor, Support Vector Machine, Phương pháp Eaton, Phương pháp Bowers

# Giới thiệu

Áp suất lỗ rỗng của hệ tầng được định nghĩa là áp suất của chất lỏng lấp đầy các lỗ của hệ tầng. Nó có thể bằng với áp suất thủy tĩnh, được gọi là áp suất lỗ rỗng thông thường, hoặc khác với áp suất thủy tĩnh (cao hơn hoặc thấp hơn), được gọi là áp suất lỗ rỗng bất thường. Swarbrick và Osborne (1998) đã mô tả một số cơ chế tạo ra áp suất bất thường như sự giãn nở thủy sản, sự mất cân bằng khi nén chặt (dưới sự nén chặt), nứt khí và khai thác hydrocacbon, sự nổi hydrocacbon, ứng suất bên do các sự kiện kiến tạo và biến đổi khoáng chất. Sự thiếu hụt trong dự đoán áp suất lỗ rỗng hình thành trước và trong quá trình khoan sẽ làm tăng khả năng xảy ra rủi ro và sự cố trong quá trình khoan. Áp suất cao bất thường gây ra sự suy yếu của đứt và núi lửa bùn .Dự đoán áp suất lỗ rỗng trong các hồ chứa cacbonat độc đáo, phụ thuộc vào các bản ghi giếng luôn là một nhiệm vụ khó khăn. Các hạt cacbonat không nén chặt đồng nhất với độ sâu như Đá phiến. Ngoài ra, hệ thống lỗ xốp trong cacbonat là sự kết hợp phức tạp của một số loại lỗ rỗng, làm tăng nguy cơ dự đoán áp suất lỗ rỗng sai.[1]

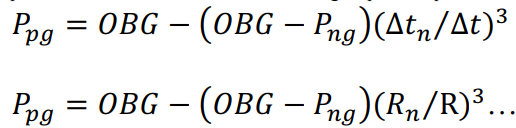
Mục tiêu của công việc này là dự đoán áp suất lỗ rỗng hình thành trong vỉa cacbonat không thông thường và một số vỉa sét, cát kết phụ thuộc vào độ rỗng và thuộc tính chịu nén của đá và bằng cách sử dụng dây wireline logging data

# Các phương pháp truyền thống và học máy

## Phương pháp truyền thống

### Phương pháp Eaton

Phương pháp Eaton (1975) là kỹ thuật được sử dụng phổ biến để dự đoán áp lực lỗ rỗng trong mỏ dầu. Nó phụ thuộc vào cả điện trở suất và dữ liệu thời gian vận chuyển nén âm trong các phương trình thực nghiệm sau:[2], [3]



OBG : ứng suất thẳng đứng

Ppg :áp suất gradient thành hệ

Png: áp suất thủy tĩnh[3]

∆𝑡𝑛 & 𝑅𝑛: thời gian truyền sống và điện trở suất chuẩn

∆𝑡 and R: thời gian truyền sống và điện trở suất

.

### Phương pháp Bowers

Bowers (1995) đã tính toán ứng suất hiệu quả từ dữ liệu áp suất lỗ rỗng đo được và ứng suất lớp phủ bên ngoài và phân tích vận tốc âm thanh tương ứng từ dữ liệu khai thác giếng ở sườn Vịnh Mexico. Ông đề xuất rằng vận tốc âm thanh và ứng suất hiệu quả có mối quan hệ năng lượng như sau:

Tại là vận tốc nén ở một độ sâu nhất định, là vận tốc nén qua mud log (thường là 5000 ft/s) ưng suất thẳng đứng, A và B là các tham số được hiệu chỉnh với vận tốc bù so với dữ liệu ứng suất hiệu quả.[2]

Ta có: và

Với A = 10-20 và B = 0.7-0.75 đối với giếng theo Gulf of Mexico

## Phương pháp học máy

### Mô hình Multiple linear Regression

Trong trường hợp đơn giản nhất với một biến đầu vào phương trình tuyến tính có dạng:

Với a,b là các hằng số,a 0

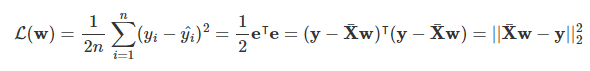
Vì có một biến đầu vào nên phương trình trên còn được gọi là phương trình hồi qui tuyến tính đơn biến. Trên mặt phẳng hai chiều, quan hệ giữa và được thể hiện là một đường thẳng.

Trong trường hợp tổng quát phương trình tuyến tính có dạng[4]

Phương trình trên có nhiều hơn một biến đầu vào nên được gọi là phương trình hồi qui tuyến tính đa biến. Tập hợp những điểm tạo thành một mặt phẳng (plane) trong trường hợp hai chiều và siêu phẳng (hyperplane) đối với không gian nhiều hơn hai chiều.

Hồi qui tuyến tính đa biến là hồi qui tuyến tính với nhiều hơn một biến đầu vào. Hồi qui tuyến tính đa biến phổ biến hơn so với đơn biến vì trên thực tế rất hiếm các tác vụ dự báo chỉ gồm một biến đầu vào. Phương trình hồi qui của nó có dạng:

Hàm mất mát MSE là trung bình tổng bình phương của các sai số nên nó có dạng:



Nghiệm của phương trình hồi qui:



Phương hình hồi qui đa biến có nghiệm khi A là khả nghịch.

### Mô hình Random Forest Regressor

. Cây quyết định là một mô hình khá nối tiếng hoạt động trên cả hai lớp bài toán phân loại và dự báo của học có giám sát. Ý tưởng chính của mô hình là xây dựng một đồ thị dạng câu hỏi để đưa ra dự báo.

Dù có độ chính xác khá cao nhưng cây quyết định tồn tại những hạn chế lớn đó là:[5]

* Dễ xảy ra quá khớp nếu số lượng các đặc trưng để hỏi lớn. Khi độ sâu của cây quyết định không bị giới hạn thì có thể tạo ra những node lá chỉ có một vài quan sát. Những kết luận dự báo từ chúng thường chỉ đúng trên tập huấn luyện mà không đúng trên tập kiểm tra.
* Trong tình huống bộ dữ liệu có số lượng biến lớn. Một cây quyết định có độ sâu giới hạn (để giảm thiểu quá khớp) thường bỏ sót những biến quan trọng.
* Cây quyết định chỉ tạo ra một kịch bản dự báo duy nhất cho mỗi một quan sát nên nếu model có hiệu suất kém thì kết quả sẽ bị chệch.

Nếu như sức mạnh của một cây quyết định là yếu thì hợp sức của nhiều cây quyết định sẽ trở nên mạnh mẽ hơn (một cây làm chẳng nên non, ba cây chụm lại nên hòn núi cao). Ý tưởng của sự hợp sức đã hình thành nên mô hình rừng cây (Random Forest).[6]

Mô hình rừng cây sẽ áp dụng cả hai phương pháp học kết hợp (ensemble learning) và lấy mẫu tái lập (boostrapping). Thứ tự của quá trình tạo thành một mô hình rừng cây như sau:[7]

* Lấy mẫu tái lập một cách ngẫu nhiên từ tập huấn luyện để tạo thành một tập dữ liệu con.
* Lựa chọn ra ngẫu nhiên d biến và xây dựng mô hình cây quyết định dựa trên những biến này và tập dữ liệu con ở bước 1. Chúng ta sẽ xây dựng nhiều cây quyết định nên bước 1 và 2 sẽ lặp lại nhiều lần.
* Thực hiện bầu cử hoặc lấy trung bình giữa các cây quyết định để đưa ra dự báo.

Kết quả dự báo từ mô hình rừng cây là sự kết hợp của nhiều cây quyết định nên chúng tận dụng được trí thông minh đám đông và giúp cải thiện độ chính xác so với chỉ sử dụng một mô hình cây quyết định.

Nếu như mô hình cây quyết định thường bị nhạy cảm với dữ liệu ngoại lai (outlier) thì mô hình rừng cây được huấn luyện trên nhiều tập dữ liệu con khác nhau, trong đó có những tập được loại bỏ dữ liệu ngoại lai, điều này giúp cho mô hình ít bị nhạy cảm với dữ liệu ngoại lai hơn.

Sự kết hợp giữa các cây quyết định giúp cho kết quả ít bị chệch và phương sai giảm. Như vậy chúng ta giảm thiểu được hiện tượng quá khớp ở mô hình rừng cây, một điều mà mô hình cây quyết định thường xuyên gặp phải.

Cuối cùng các bộ dữ liệu được sử dụng từ những cây quyết định đều xuất phát từ dữ liệu huấn luyện nên quy luật học được giữa các cây quyết định sẽ gần tương tự như nhau và tổng hợp kết quả giữa chúng không có xu hướng bị chệch

### Mô hình Support Vector Machine

SVM là viết tắt của cụm từ *support vector machine*. Đây là một thuật toán khá hiệu quả trong lớp các bài toán phân loại nhị phân và dự báo của học có giám sát. Thuật toán này có ưu điểm là hoạt động tốt đối với những mẫu dữ liệu có kích thước lớn và thường mang lại kết quả vượt trội so với lớp các thuật toán khác trong học có giám sát.

Ưu điểm của SVM đó là:[4]

* Đây là thuật toán hoạt động hiệu quả với không gian cao chiều (*high dimensional spaces*).
* Thuật toán tiêu tốn ít bộ nhớ vì chỉ sử dụng các điểm trong *tập hỗ trợ* để dự báo trong *hàm quyết định*.
* Chúng ta có thể tạo ra nhiều *hàm quyết định* từ những hàm kernel khác nhau. Thậm chí sử dụng đúng kernel có thể giúp cải thiện thuật toán lên đáng kể.

Chính vì tính hiệu quả mà SVM thường được áp dụng nhiều trong các tác vụ phân loại và dự báo, cũng như được nhiều công ty ứng dụng và triển khai trên môi trường production. Chúng ta có thể liệt kê một số ứng dụng của thuật toán SVM đó là:

* Mô hình chuẩn đoán bệnh. Dựa vào biến mục tiêu là những chỉ số xét nghiệm lâm sàng, thuật toán đưa ra dự báo về một số bệnh như tiểu đường, suy thận, máu nhiễm mỡ,…
* Trước khi thuật toán CNN và Deep Learning bùng nổ thì SVM là lớp mô hình cực kì phổ biến trong phân loại ảnh.
* Mô hình phân loại tin tức. Xác định chủ đề của một đoạn văn bản, phân loại cảm xúc văn bản, phân loại thư rác.

Mô hình phát hiện gian lận[8], [9]

# Kết quả và thảo luận

## Dữ liệu tính toán

Đầu vào: độ sâu (depth), Grammray (GR), log mật độ (RHOB), tốc độ song sonic (Vp), tốc độ song chuyền qua shale (Vsh), đường kính (Caplier),độ rỗng (Porosity),điện trở suất ( Resistivity), ứng suất (stress), áp suất lỗ rỗng được ghi nhận (Pp)

Đầu ra: áp suất lỗ rỗng tính toán bằng phương pháp Eaton, Phương pháp Ratio, Phương pháp Bowers, các mô hình machine learning (Multiple Linear Regression, Random Forest,SVM.

## Tính toán bằng phương pháp Eaton

* Tính toán áp suất thủy tĩnh (Png)

Công thức tổng quát:

Khối lượng của nước:

Gia tốc trọng trường:

Độ sâu:

Ta có :

MegaPascal (MPA) = Pascal = 10 Bars = 145.037738 Psi

Pascal = 0.00014503778 Psi

Vậy ta được phương trình sau khi biến đổi :

* Tính ứng suất thẳng đứng (Overburden Stress)

Ứng suất thẳng đứng () là áp lực tác dụng lên thành hệ ở một độ sâu nhất định do tổng trọng lượng của các loại đá và các chất lỏng theo chiều sâu.

σv có thể được xác định bằng tích phân khối lượng riêng tổng của thành hệ từ dữ liệu log bằng wire-line, theo công thức:

Với:

là chiều sâu mực nước,

là khối lượng riêng của nước biển,

Z là chiều sâu thẳng đứng của giếng,

ρb là khối lượng riêng tổng của thành hệ.

Ứng suất thẳng đứng luôn có phương thẳng đứng, ứng suất thẳng đứng có ý nghĩa hết sức quan trọng vì thông thường nó đóng vai trò cực đại ở những độ sâu nghiên cứu lớn khoảng từ 2000m.[10]

Nguồn gây ứng suất thẳng đứng cũng có thể từ điều kiện địa chất khác như xâm nhập magma và các vòm muối xung quanh khu vực thành hệ đá. Dựa vào mối quan hệ đó người ta có thể tính được σv bằng các phương pháp phân tích đường log mật độ khối (RHOB), phân tích mật độ mẫu lõi.

Phương pháp thứ hai dùng để xác định σv là sử dụng tài liệu địa chấn. Cụ thể là dùng phương pháp Gardner:

Trong đó:

ρ là mật độ khối,

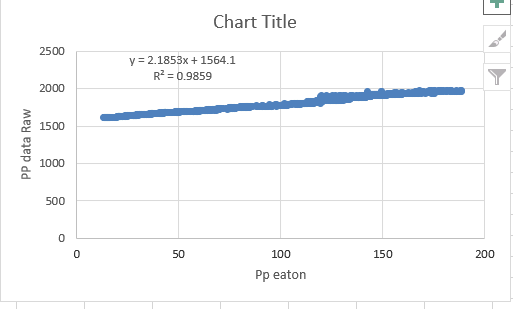
Vp là sóng cắt,

a, b là hằng số, thường lấy a = 0.23 và b = 0.25.

Ta có công thứ sau =

* Tính ứng suất thẳng đứng gradient (Overburden stress Gradient)[11]

Overburden stress Gradient



Hình 1: Kết quả tính toán áp suất lỗ rỗng theo phương pháp Eaton:

## Phương pháp Bowers

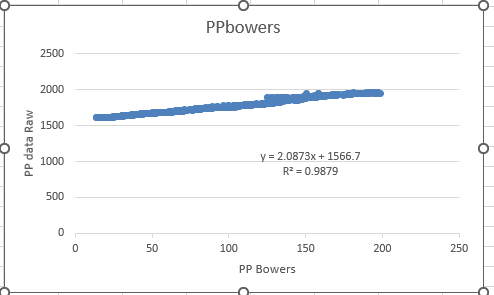
Ta có mối liên hệ giữa vận tốc truyền âm và ứng suất hiệu dụng:

: là vận tốc song âm truyền ở độ sâu nhất định

: vận tốc truyền trong mùn khoan = 5000ft/s hay 1520 m/s

A và B là tham số hiệu chình vấn tốc bì trừ với hiệu chỉnh dữ liệu ứng suất hiệu dụng

Ta có: và và



Hình 2: Kết quả tính toán áp suất lỗ rỗng theo phương pháp Bowers:

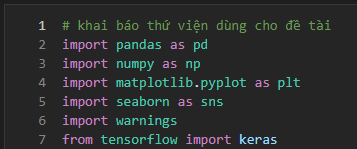
## Ứng dụng mô hình học máy cho việc dự đoán áp suất lỗ rỗng

### Xây dựng mô hình

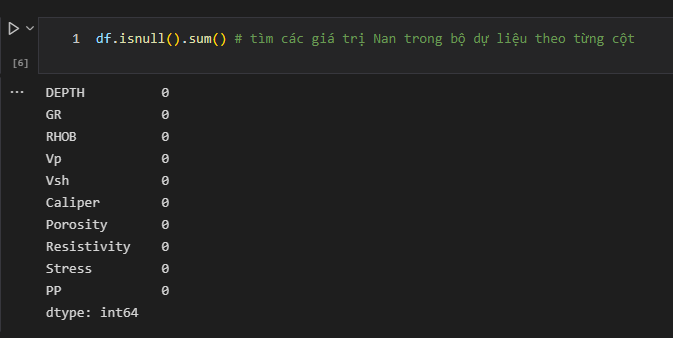
Từ tập dữ liệu tổng ban đầu 7002 điểm dữ liệu để tăng độ tin cậy cho bộ dữ liệu

Tập con để train (80%): đây là tập dữ liệu để chúng ta xây dựng mô hình mẫu

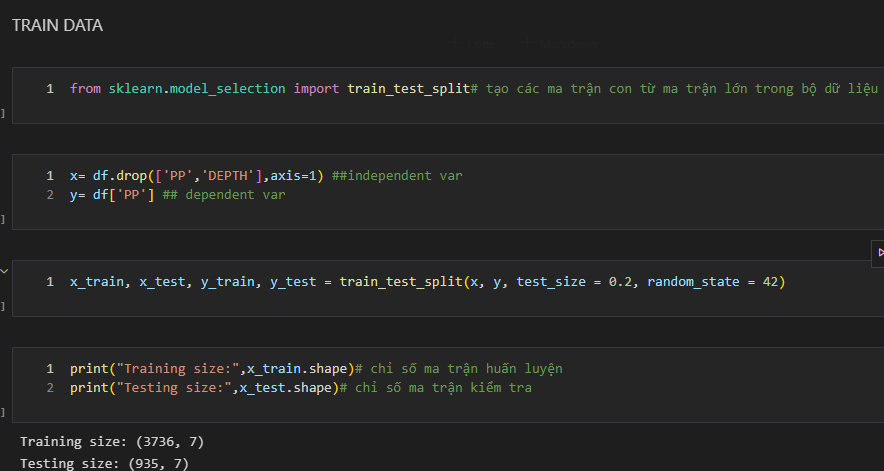
Tập con để test (20%): tập dữ liệu test dùng để so sánh với tập train đánh giá độ chính xác của tập train.



Hình 3 các thư viện sử dụng để xử lý dữ liệu các mô hình



Hình 4 xác dịnh các điểm dữ liệu trống trong bộ dữ liệu



Hình 5: xây dựng mô hình mẫu

### Mô hình Multiple linear Regression

Ta sử dụng thư viện: from sklearn.linear\_model import LinearRegression

Ta huấn luyện mô hình Multiple Linear Regression

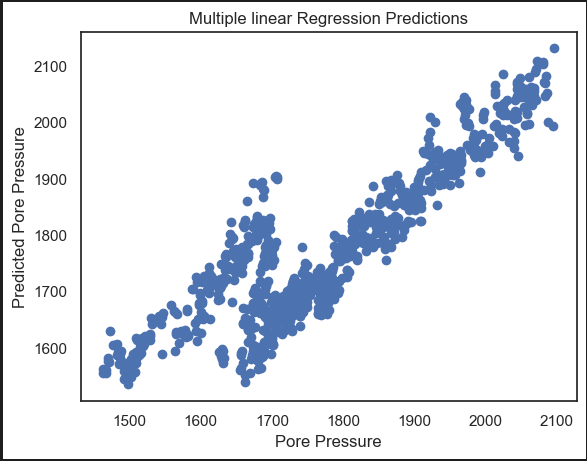
model = LinearRegression(fit\_intercept= True, normalize=True)

model.fit(x\_train,y\_train)

pred=model.predict(x\_test)



Hình 6 mô hình Multiple Linear Regression



Hình 7 kết quả áp suất lỗ rỗng sau khi dùng mô hình Multiple Linear Regression

### Mô hình Random Forest Regressor

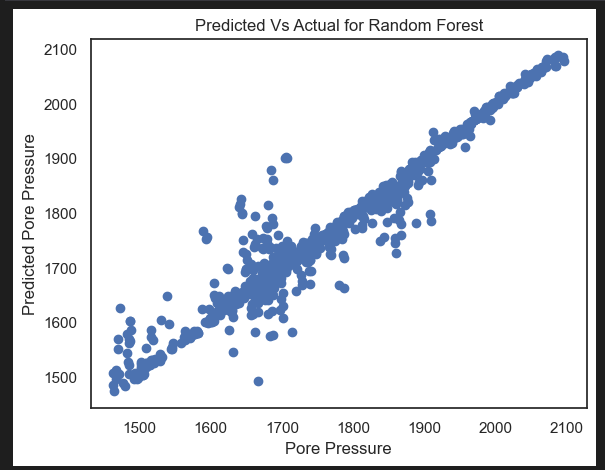
Ta dùng thư viện from sklearn.ensemble import RandomForestRegressor

Ta huấn luyện mô hình :

reg=RandomForestRegressor()

reg.fit(x\_train, y\_train)

pred=reg.predict(x\_test)



Hình 8 kết quả áp suất lỗ rỗng khi sử dụng mô hình Random Forest Regressor

### Mô hình Support Vector Machine

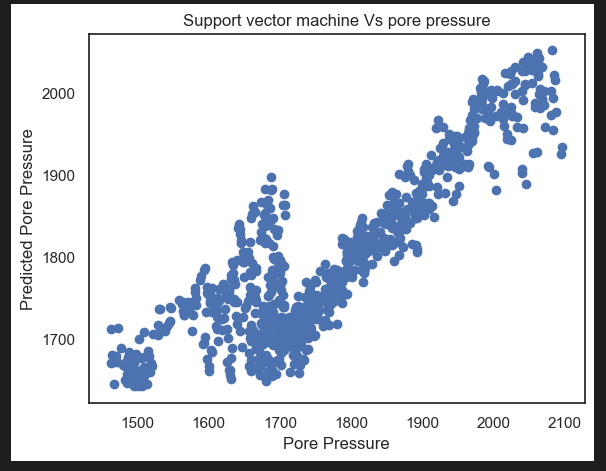
Ta dùng thư viện from sklearn import svm

Ta huấn luyện mô hình

sv = svm.SVR()

sv.fit(x\_train,y\_train)

pred1=sv.predict(x\_test)



Hình 9 kết quả áp suất lỗ rỗng khi sử dụng mô hình Support Vector Machine

## So sánh và đánh giá các phương pháp tính toán độ rỗng

Để đánh giá sự chính xác của các phương pháp dự đoán độ rỗng ở trên tác giả tiến hành so sánh các sai số của từng phương pháp (truyền thống, Multiple Linear Regression, Random Forest Regressor, Support Vector Machine). Các thông số độ lệch thống kê được sử dụng bao gồm: sai số toàn phương trung bình (MSE – Mean squared error), căn bậc hai của sai số toàn phương trung bình (RMSE – Root mean squared error), hệ số xác định (R2 – Coefficient of determination).

Giá trị của các sai số thống kê và hệ số tương quan của các phương pháp được thể hiện trong Bảng 1

**Bảng 1:** So sánh sai số tính toán độ rỗng

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| STT | Phương pháp | MSE | RMSE | R2 |
| 1 | MLP | 4805.267109 | 0.693200 | 78.61968 |
| 2 | RF | 1296.571285441739 | 0.360079 | 94.272555 |
| 3 | SVM | 6768.838659177967 | 0.822729 | 69.883061 |

# Kết luận

Với việc tính toán và so sánh các phương pháp khác nhau, nghiên cứu đã chỉ ra những điểm ưu việt của BA phương pháp học máy Multiple Linear Regression, Random Forest Regressor, Support Vector Machine so với các phương pháp truyền thống. Từ kết quả thu được, cả hai phương pháp học máy đều cho kết quả chính xác, nhanh chóng và tiện lợi hơn việc tính toán bằng phương pháp truyền thống.

# Tài liệu tham khảo

[1] \* Exxon Production Research Co Glenn L. Bowers, “Pore Pressure Estimation From Velocity Data: Accounting for Overpressure Mechanisms Besides Undercompaction,” 1995.

[2] M. Elmahdy, E. Farag, E. Tarabees, and A. Bakr, “Pore Pressure Prediction in Unconventional Carbonate Reservoir,” *Society of Petroleum Engineers*, 2018.

[3] Glenn L. Bowers, “Determining an Appropriate Pore-Pressure Estimation Strategy,” *Offshore Technology Conference*, 2021.

[4] Miguel Angel Alvarez, “Pore pressure prediction from well logs: methods, modifications, and new approaches,” *Shell Exploration and Production Company*, 2011.

[5] P S Hutomo, M S Rosid, and M W Haidar, “Pore Pressure Prediction Using Eaton and Neural Network Method in Carbonate Field ‘X’ Based on Seismic Data,” *Materials Science and Engineering*, 2019.

[6] Reza Majidi and Martin Albertin, “Pore-Pressure Estimation by Use of Mechanical Specific Energy and Drilling Efficiency,” *SPE Drilling & Completion*, 2017.

[7] Honglin Huang *et al.*, “Research on prediction methods of formation pore pressure based on machine learning,” *Energy Sci Eng*, 2022.

[8] PETER B. FLEMINGD, *A CONCISE GUIDE TO GEOPRESSURE*. 2002.

[9] G.V. CHILINGAR and V.A. SEREBRYAKOV, *ORIGIN AND PREDICTION OF ABNORMAL FORMATION PRESSURES*. 2002.

[10] Zahra Bahmaei and Erfan Hossein, “Pore pressure prediction using seismic velocity modeling: case study, Sefid‑Zakhor gas field in Southern Iran,” *J Pet Explor Prod Technol*, 2019.

[11] A. Ogbamikhumi, O. M. Hamid-Osazuwa, and E. A. Imoru, “Post Drill Pore Pressure Prediction for Geo-hazard Assessment of Offset Wells in Hamoru Field,” *POST DRILL PORE PRESSURE PREDICTION FOR GEO-HAZARD ASSESSMENT OF WELLS*, 2020.