Machine Learning Algorithms-Based Classification of Lithology using Geophysical Logs: ICDP DSeis Project Boreholes, South Africa.

By

Obehi Chapet Atita

Supervisor(s):

Prof Raymond Durrheim and Dr Eric Saffou

A Dissertation submitted to the

Faculty of Science, University of the Witwatersrand, Johannesburg,

in Fulfilment of the requirements for the degree of

Master of Science

School of Geosciences

University of the Witwatersrand

Johannesburg, 2024

Declaration

I declare that this dissertation is my own, unaided work. It is being submitted for the degree of Master of Science at the University of the Witwatersrand, Johannesburg. It has not been submitted before for any degree or examination in any other University.

Obehi Chapet Atita

2nd day of September 2024 at the University of the Witwatersrand, Johannesburg.

ABSTRACT

One of the most significant geosciences tasks is the accurate classification of lithologies for metal and mineral resources exploration, characterization of oil/gas reservoir(s), and the planning and management of mining operations. With the availability of abundant, huge and multidimensional datasets, machine learning-based data-driven methods have been widely adopted to assist in solving geoscientific problems such as the efficient evaluation and interpretation of large datasets. The adoption of machine learning-based methods aims to improve lithological identification accuracy and extract information required for accurate and objective decision-making with respect to activities such as exploration, drilling, mine planning and production. Practically, this helps to reduce working time and operating costs.

We aim to evaluate the feasibility of machine learning-based algorithms application to geophysical log data for the automated classification of lithologies based on the stratigraphic unit at the formation level for the purpose of distinguishing and correlating the quartzites between boreholes, and mapping key radioactive zones within the mining horizon. This study implemented four different machine learning algorithms: gradient boosting decision trees, random forest, support vector machine, and K-means clustering models. Analyzed features and labelled datasets are multivariate downhole geophysical and lithology logs from the two ICDP DSeis project boreholes drilled in the Klerksdorp gold field, respectively. To mitigate misclassification error and avoid model overfitting/underfitting, the optimal combination sets and optimal values for each implemented supervised model's hyperparameters were obtained using the Grid search and 10-fold cross-validation optimization methods. The input dataset was randomly split automatedly into training and testing subsets that made up 80% and 20% of the original dataset, respectively. The models were trained and cross-validated using the training subset, and their performances were assessed using the testing subset.

The classification performance of each model was evaluated using F1 scores and visualized using confusion matrices. The best supervised classification model for our study area was selected based on the testing subset F1 scores and computational cost of training models. The testing subset results shows that Random Forest and Support Vector Machine classifier models performed much better

relative to the Gradient Boosting Decision Trees classifier model, with F1 scores over 0.80 in borehole A and B. In borehole A and B, Random Forest classifier has the least computational training time of about 14- and 6- hours, respectively. The feature importance results demonstrate that the logging feature P-wave velocity (V_p) is the highest predicting feature to the lithology classification in both boreholes. We find that the quartzite classes at different stratigraphic positions in each borehole are similar and they are correlated between the DSeis boreholes. The K-means clustering revealed three clusters in this study area and effectively map the radioactive zones.

This study illustrates that geophysical log data and machine learning-based algorithms can improve the task of data analysis in the geosciences with accurate, reproducible and automated prediction of lithologies, correlation and mapping of radioactive zones in gold mine. This study outputs can serve as quality control measures for future similar studies both in the academic and industry. We identified that availability of large data is the major factor to high accuracy performance of machine learning-based algorithms for classification problems.

ACKNOWLEDGEMENTS

To God almighty, thank you so much for giving me the strength to carry on till the finish line.

I would like to thank my supervisors, Prof Raymond Durrheim and Dr. Eric Saffou for their support, guidance and enthusiasm to review my work and provide me with feedback. An extra thanks to Prof Raymond Durrheim for going beyond his normal responsible to financially support me.

1wish to express my appreciation to DSeis project and Prof Hiroshi Ogasawara for availing the geological and geophysical logs in the Moab Khotsong gold mine. A huge thanks to Prof Glen Nwaila for always being available to review my work and provide feedback despite his busy schedule. I would also like to thank Ndamulelo Mutshafa for making his office and computers available to conduct my research.

I express my greatest appreciation to Agate-project for the financial support I received toward the 2nd year of my MSc.

To my family, thank you for always encouraging me to get my work done, when I had no strength to carry on. I would like to thank my friends at the postgraduate office Senamile Dumisa, Desiderius Chapewa, Berhanu Robel, and Joshua Pillay for accommodating my inconveniences during my most financial struggle times. A very big thank you to Sikelela Gomo and Emmanuel Onyebueke for always checking up on me and motivating me to finish my research.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABBREVIATIONS

ML: Local Magnitude

ICDP: International Continental Scientific Drilling Program

DSeis: Drilling into Seismogenic Zones in deep South Africa Gold Mines

V_p: P-wave velocity

V_s: S-wave velocity

SNR: signal-to-noise ratio

UR: Upper Roodepoort

LR: Lower Roodepoort

RF: Random Forest

SVM: Support Vector Machine

GBDT: Gradient Boosting Decision Tree

IMS: Institute of Mine Seismology

SATREPS: Science and Technology Research Partnership for Sustainable Development

1. INTRODUCTION

Good geological knowledge, particularly of lithologies and their spatial variations, is important for various aspects of mining, including identification of mineral deposits, detection of areas with high mineral exploration potential, mine planning and modelling, and hazard mitigation. Apart from the importance of providing subsurface rock information for objective decision making concerning mine planning and modelling and mineral resource exploration, detailed understanding of lithology types and properties has practical importance when performing ore body reserve and mineral resource evaluations at the exploration stage and optimally exploiting the resources at the development stage (Salehi and Honarvar, 2014; Xie et al., 2018). This is because lithology controls physical property distributions (e.g., density, porosity, velocities, and permeability) and transport mechanisms of sediments and ore-forming elements (Pan et al., 2008; Li and Anderson-Sprecher, 2006; Sun et al., 2019; Wang et al., 2018; Lai et al., 2013; Ren et al., 2019).

Identification of subsurface formation lithological units was traditionally conducted by applying knowledge-driven methods such as the observation and analysis of borehole core samples, and inference from cuttings obtained during drilling operations. These traditional techniques are most practical and direct, but reliant on the expertise of geologists, leading to varying results and uncertainties (Akinyokun et al., 2009; Xie et al., 2018; Zou et al., 2021). In addition, coring is costly, and hence sometimes limited to targeted borehole sections only. Core recovery can be costly and sometimes incomplete (it may even be less than 50%) or, in the case of percussion drilling, where cores are not recovered at all (Benaouda et al., 1999; Li and Anderson-Sprecher, 2006; Mahmoodi et al., 2016; Zhang et al., 2017; Ren et al., 2019; Zou et al., 2021), prompting the adoption of geophysical log data for improved lithology identification (Hsieh et al., 2005; Jahdhami and Anboori, 2017; Bressan et al., 2020).

Geophysical log data offers advantages such as high vertical resolution, good continuity of *in-situ* data significant for mapping subsurface geological structures, and acquisition convenience compared to traditional lithology identification methods (Xie et al., 2018). Also, downhole geophysical log dataset identifies parts of the borehole with no core recovery (Benaouda et al., 1999). Several graphical methods have been implemented for lithology identification using geophysical log dataset. These includes histograms, cross plots (McDowell et al., 1998; Killeen,

1997; Ohakwere-Eze et al., 2018; Ren et al., 2019), and reconstruction of logging feature characteristic curve from many logs and logging shape description (Guoming, 2010; Lai et al., 2013; Zhang et al., 2014; Guiwen et al., 2015; Ren et al., 2019). However, the complex geological conditions and non-linear relationship between the geophysical logging signatures and subsurface lithology sequence (i.e., there is no defined mathematical method to quantify the relationship between the information denoted by logs and the real subsurface lithology sequence) poses limitations to the use of linear logging signature equations and empirical statistical formulas, especially in hard rock environments where the signatures of geophysical log curves for some different lithological types overlap, mostly at the transition zone (Li and Anderson-Sprecher, 2006; Mahmoodi et al., 2016; Ren et al., 2019).

In recent years, machine learning-based algorithms have gained widespread acceptance in the geosciences for lithology prediction due to their ability to handle complex data relationships and improve prediction accuracy. Several studies have demonstrated the effectiveness of machine learning-based algorithms in lithology prediction, showing superior performance compared to conventional methods. Zhang et al. (2021) demonstrated that machine learning-based algorithms can accurately and robustly predict Archaean shales, a task challenging for conventional methods due to their variability. They noted that the accuracy of these machine learning-based algorithms is not affected by chemical variability and geological complexity but is strongly influenced by samples size, as accuracy increases with increasing sample size.

Similarly, Zou et al. (2021) showed that gradient boosting decision tree model outperforms artificial neural network, support vector machine, random forest, and AdaBoost models, in classifying typical lithology associated with mineral deposits using well log data. In studies by Xie et al. (2018); Sun et al. (2019); Bressan et al. (2020); and Merembayev et al. (2021), it was concluded that random forest algorithm was the most suitable classifier among algorithms they considered for lithology classification using well log data such as gamma ray, caliper, spontaneous potential, thorium, and density in production applications. This conclusion was drawn based on its higher accuracy, lower prediction errors, and faster model training time compared to other algorithms. Furthermore, Rodriguez-Galiano et al. (2015) investigated the capability of machine learning regression algorithms in predictive mapping of epithermal gold mineral prospectivity

using geological, geochemical, and geophysical data. They found that random forest model performed best in terms of mapping accuracy, parameterization need, and sensitivity to the training sample size.

The widespread adoption of machine learning algorithms in the mining industry and academia community relies on the availability of sufficient high-quality datasets and user-friendly software packages with pre-defined machine learning analysis procedures. Geophysical logging is one technique that provides abundant high-quality datasets for the application of machine learning-based algorithms in the geosciences (Zhang et al., 2017), offering opportunities to enhance lithology identification and geological understanding. Machine learning-based algorithms offer more objective and data-driven practical guidance for future sampling, mine operation planning and mineral exploration by effectively tackling the non-linear relationship between logging signatures and formation lithologies, reducing reliance on experience and knowledge of geoscientists (Pechnig et al., 2005; Konate et al., 2015; Mahmoodi and Smith, 2015; Han et al., 2018; and Zou et al., 2021). Machine learning-based algorithms were proposed to be more efficient and reliable for identification of lithology by providing non-linear models that project lithology identification as a case of classification (Zou et al., 2021). The classification of lithology applying geophysical logs is based on the established relationship between the inferred lithology and the geophysical logging signatures (Sebtosheikh and Salehi, 2015; Ren et al., 2019; Min et al., 2020; and Zou et al., 2021).

1.1. Aim and Objectives

This study aims to assess the feasibility of implementing optimized machine learning-based algorithms including Random Forest, gradient boosting decision trees, and support vector machine models, for lithology classification using a suite of downhole geophysical and lithological logs. Specifically, the study seeks to demonstrate the effectiveness of these algorithms in accurately distinguishing the most dominant rock class (quartzite) at different stratigraphic levels.

Additionally, the study aims to demonstrate the ability of mapping key radioactive zones using K-means clustering analysis solely based on the combination of geophysical log data (total gamma ray, density, V_p and V_s). The random forest, gradient boosting decision trees, and support vector

machines algorithms were chosen because they are well-known techniques widely applied in data analysis procedures to identify formation lithology in the geosciences due to their high performance characteristics in lithology classification and robustness to prediction bias, outliers, imbalanced data, and small datasets (e.g., Xie et al., 2018; Sun et al., 2019; Zhang et al., 2021; Zou et al., 2021; Nwaila et al., 2022). In addition, K-means clustering of geophysical logs (density, V_p, V_s, and gamma ray) may create meaningful clusters that can be associated with known existing lithological types, which can be applied in further prediction of similar lithology type using geophysical log data. Also, clustering of geophysical log data may be used to refine or validate existing geological logs.

Despite the availability of geophysical log data from the ICDP DSeis boreholes, no research has utilized data-driven methods, such as machine learning algorithms for lithology classification in this context. To address this gap, our study seeks to leverage the advantages of these machine learning algorithms for accurate lithological classification and mapping to validate existing suggested traditional lithology classification.

The following steps will be carried out to achieve our aim:

- Build three optimized supervised machine learning-base models (gradient boosting decision trees, Random Forest and support vector machine) using a training subset (80% of the original dataset),
- Compare model (gradient boosting decision trees, Random Forest and support vector machine) classification performances based on their computation training time and prediction F1 scores using the testing dataset (20% of the original dataset),
- Demonstrate the impact of the correlation between downhole geophysical log features within borehole A on model classification performance,
- Correlate quartzite strata between DSeis boreholes. Quartzite strata is only targeted due to its large sample size, and
- Develop a K-means clustering model to classify lithology for the purpose of mapping key radioactive zones combing gamma ray, density, V_p and V_s logs.

1.2. Key Questions

This research work will address the following scientific questions:

- Are the most dominant lithology classes (quartzite) with larger sample size at different formation levels of DSeis boreholes at Moab Khotsong gold mine the same, based on geophysical logs (gamma ray, density, V_p and V_s)?
- Does the correlation between geophysical log features within a borehole impact machine learning classification performance in that borehole?
- Can the K-means clustering algorithm determine the appropriate number of lithology clusters to map lithologies that are radioactive in addition to identifying the radioactive zones in mineral deposits using the combination of gamma ray, density, V_p and V_s logs?

To develop supervised classification models for the sole purpose of distinguishing the different units of quartzite and intrusives, we employed six major stages of the machine learning workflow modified after Zhang et al. (2021). This includes:

1) Data preprocessing to ensure the quality of data applied, which includes dealing with missing data and outliers, and standardization,
2) Data splitting into training and testing datasets to build and evaluate the performance of models, respectively,
3) Optimization and hyperparameter tuning to ensure optimal values of models' hyperparameters are selection to avoid overfitting and prediction bias,
4) Model building for classification of lithologies using the training dataset,
5) Prediction performance evaluation using F1 score, and
6) Analysis and verification of predicted lithologies by comparing them with lithologies predicted traditionally .

For the implementation of K-means clustering analysis, two major machine learning methods were employed:

1) Data preprocessing

2) Estimation of optimal number of K using “elbow” method, assuming information about the lithological log is absent.

1.3. Study Area and Project Background

The research region (Moab Khotsong gold mine; Error! Reference source not found.) is situated in the Klerksdorp goldfield region, about 170 km WSW of Johannesburg close to Orkney and Klerksdorp towns in South Africa (Mngadi, 2015; Midzi et al., 2015; Berset, 2017). The Klerksdorp goldfield is in the Witwatersrand Basin’s northwest sector. Moab Khotsong gold mine is among the youngest South African deep gold mines. It lies south of the Vaal River (Error! Reference source not found.) and started operation in 2003.

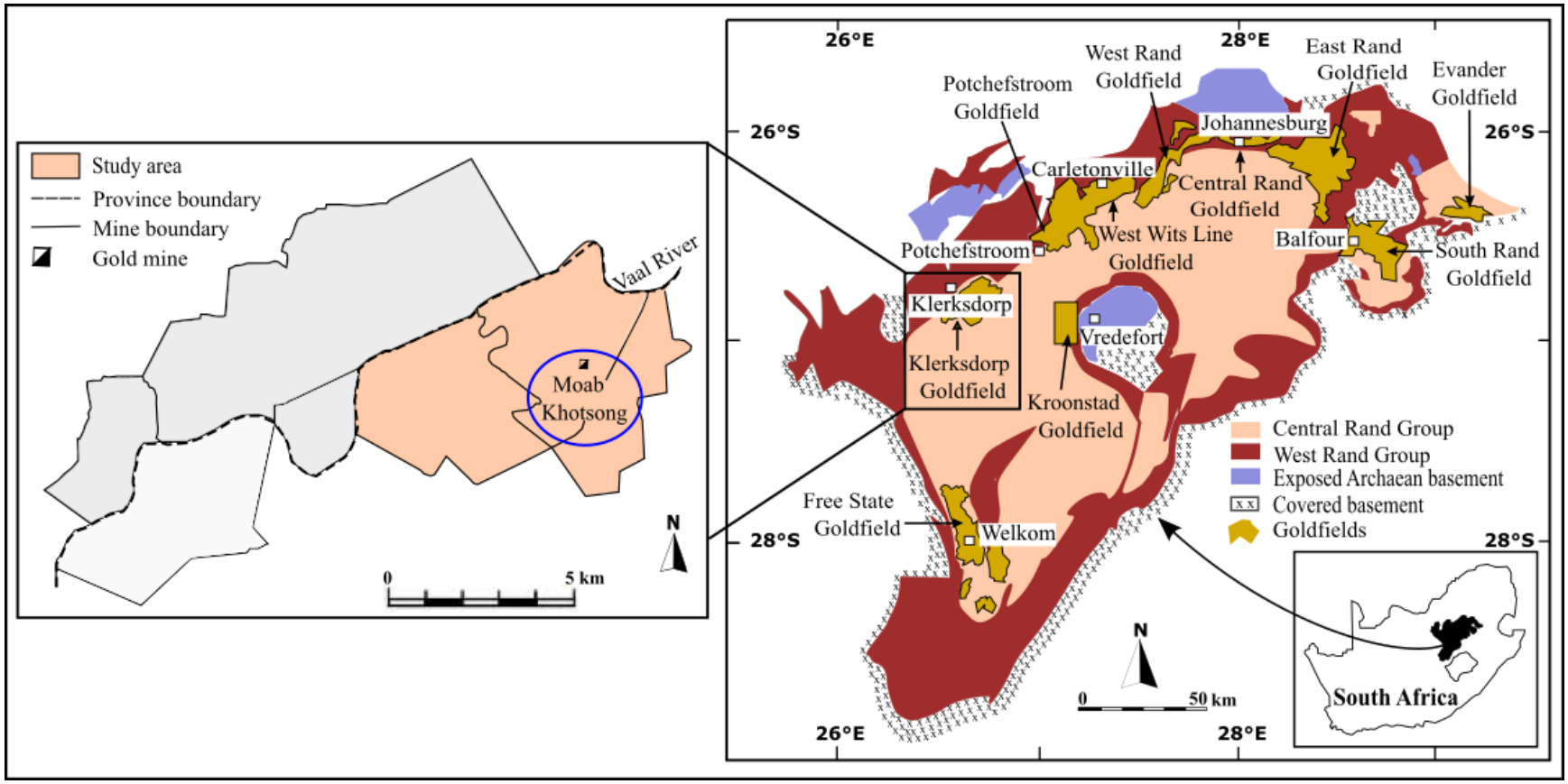

Figure 1.1: Simplified geological map of the Witwatersrand Basin showing the location of Klerksdorp goldfield. The location of Moab Khotsong mine is denoted by the blue circle (modified after Catuneanu and Biddulph, 2001; Dankert and Hein, 2010).

The gold-bearing quartz-pebble conglomerate reefs known as the Vaal and C reefs in the Central Rand Group of the late Archaean Witwatersrand Supergroup are being exploited in the Moab Khotsong gold mine (Robb and Meyer, 1995; Catuneanu and Biddulph, 2001; Dankert and Hein, 2010; Frimmel and Minter, 2002; Miyamoto et al., 2022). The mine operates at subsurface depths between 2.6- and 3.8- km, with the location of C Reef about 255 m above the Vaal Reef. The strata dip towards the southeast and northwest (Ogasawara et al., 2017; Nkosi et al., 2022). Moab Khotsong gold mine uses the scattered mining technique with a backfill system. Bracket pillars are used as stabilizing pillars to reduce the chance of slip movements on fault and dyke structures in the mining area. This is to mitigate the risks of mining-induce seismicity and rock-bursts (Berset, 2017).

1.3.1. Mining-Induced Earthquakes

The occurrence of high levels of seismicity in deep mines, typically between depths of 1.0- and 4-km, is attributed to stress alteration induced by intensive underground mining (McGarr et al. (2002). These induced earthquakes are seismic events that results from significant changes in the Earth's crust, such as an alteration in shear and normal stresses acting on a pre-existing fault plane, triggered by human activities like deep mining, quarrying, reservoir impoundment, and liquid injection at depth.

South Africa is home to the deepest gold mines in the world (Durrheim, 2015). In 1886, gold-bearing quartz pebble conglomerate reefs were found close to present-day Johannesburg. The reefs dip towards the center of an Archaean sedimentary basin that is largely overlaid by younger strata (Durrheim and Riemer, 2016; Ogasawara et al., 2017). Early in the 1890s, South Africa experienced its first mining-induced earthquake, when "extensive stopes supported solely by small reef pillars had reached depths of several hundred metres" (Durrheim and Riemer, 2016; Ogasawara et al., 2017). The 1890s mining-induced seismicity includes the Welkom earthquake (M_L 4.2), which caused the death of two miners (Uzoegbo and Li, 2002).

As modern gold mining in South African approaches depths of 4 km, the stress magnitude and temperature of rocks around the mine's workings increases and the mine operators face challenges of cooling and ventilating the workings, managing mining-induced seismicity and mitigating the risk posed by rock-bursts (Ogasawara et al., 2017).

Despite numerous advances in technology, mining-induced earthquake and rock-bursts (mining-induced earthquake is associated with rupture along weak plane of geological structures such as faults, dykes, while rock-bursts are associated with collapsing of volume of rock masses) remain one of the most serious and complex challenges of deep level mining (Durrheim and Riemer, 2016). These events pose a significant risk to mine infrastructure, miners and the public. However, they also offer a rare opportunity to study the fundamental physics of earthquakes and efforts to attempt to reliably predict the place, time and size of earthquakes by taking measurement of changes in seismic velocity induced by stress, variations in strain and tilt, acoustic and electromagnetic emissions (Durrheim, 2015).

1.3.2. The Orkney Earthquake

On 5 August 2014, an earthquake of magnitude M_L 5.5 occurred below the Moab Khotsong gold mine near Orkney in the Klerksdorp goldfield at a depth of about 4.7 ± 1.2 km below surface (Midzi et al., 2015) as seen in Error! Reference source not found.**b**. Despite the fact that in South African deep gold mines, rock-bursts and earthquakes caused by mining activities are frequent, this M_L 5.5 event was the largest ever recorded in a mining area (Midzi et al., 2015; Ogasawara et al., 2017; Manzunzu et al., 2017).

Several buildings were damaged in neighboring towns such as Orkney, Stilfontein and Khuma situated at about 16-, 12-, and 11- km away from the epicentre, respectively (Midzi et al., 2015). The mainshock was felt as far as 600 km away from the earthquake epicentre (for example, as far as Johannesburg, Pretoria, Durban, Gaborone in Botswana and Maputo in Mozambique). However, there was no significant damage to underground infrastructure, and a falling wall killed one person in Kanana, North west Province (Midzi et al., 2015; Ogasawara et al., 2017; Manzunzu et al., 2017; Ogasawara et al., 2018). According to information from the news media (News24, 2014; BBCNews, 2014), the operating company (AngloGold Ashanti) stated that 17 workers at two mines in the Orkney area suffered minor injuries (Midzi et al., 2015).

A well instrumented network of 17 surface strong-motion seismographs operated by the Council for Geoscience (CGS), 46 in-mine geophones at 2-3 km depths operated by Institute of Mine Seismology (IMS) on behalf of the mine owner, and three Ishii-type strain meters at 2.8- and 2.9-

km depths operated by Science and Technology Research Partnership for Sustainable Development (SATREPS), all recorded the mainshock and more than 800 aftershocks in 15 days after the main event were recorded. About 400 of the aftershocks were recorded during the first 24 hours after the event. The mainshock was estimated from the analysis of the seismograms to have happened at a depth of 4.7 ±1.2 km below the surface (Midzi et al., 2015; Manzunzu et al., 2017). The aftershocks were grouped into a cloud consisting of two clusters. The first cluster was located below surface at a depth of about 1-2 km on the middle mining horizon of the Moab Khotsong gold mine. The second cluster, located at a depth between 2.5- and 5- km below surface, formed the active seismogenic fault zone. They were arranged on a distinct sub-vertical geological structure of about 18 km in length (Error! Reference source not found.) with left-lateral strike-slip faulting mechanisms striking NNW-SSE (Manzunzu et al., 2017; Ogasawara et al., 2017; Ogasawara et al., 2018; Miyamoto et al., 2022). The Moab Khotsong gold mine is the nearest deep gold mine to the Orkney active zone (Ogasawara et al., 2017; Ogasawara et al., 2018; Miyamoto et al., 2022).

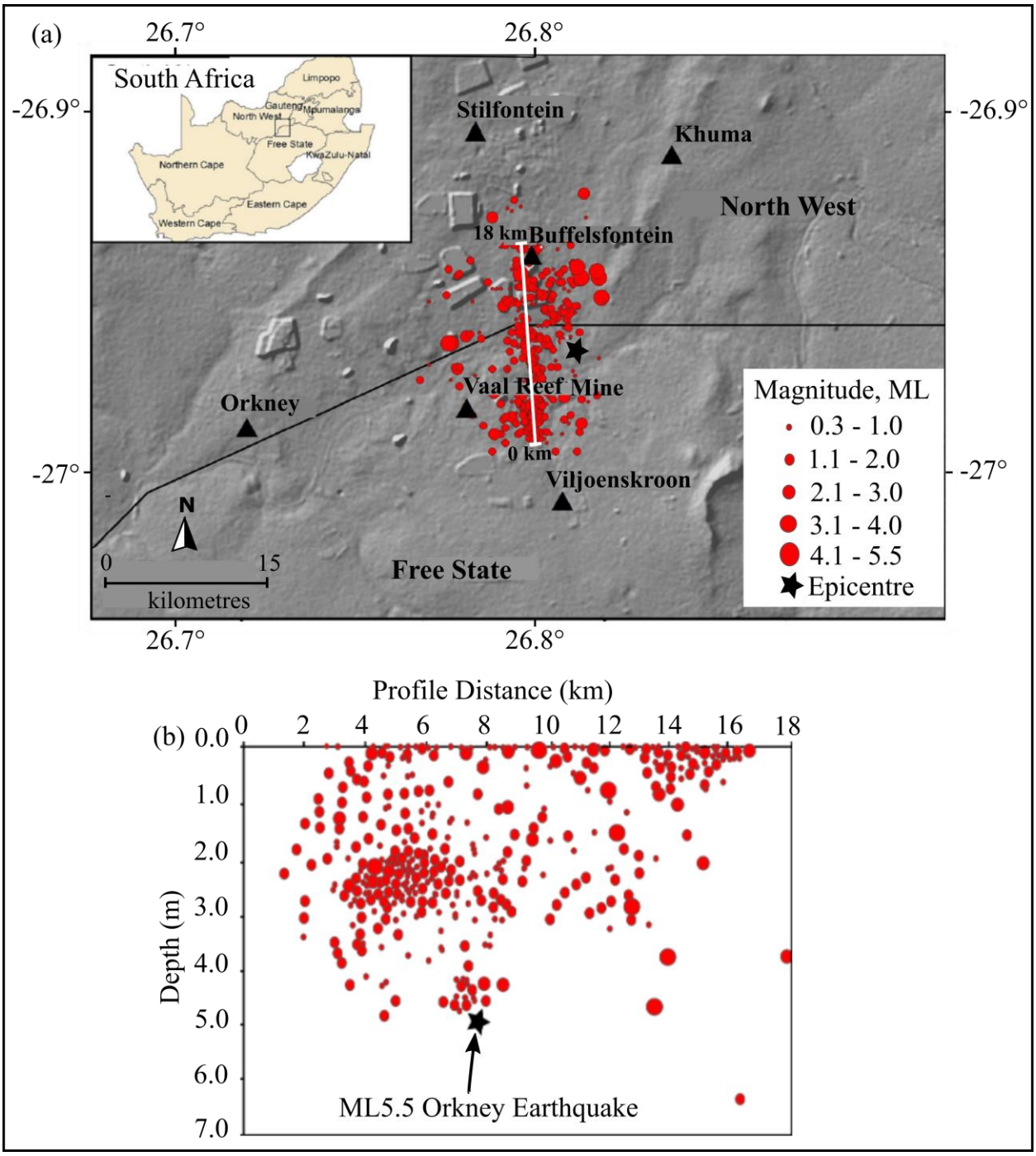

Figure 1.2: The 2014 Orkney event showing the: (a) Mainshock (black star) with distribution of associated aftershocks (red points) and its cloud profile distance (white line) and (b) Aftershock distribution according to the profile distance in a (modified after Midzi et al., 2015).

1.3.3. ICDP DSeis Drilling

Following the 5 August 2014 Orkney earthquake, a project proposal was made to the International Continental scientific Drilling Program (ICDP) in January 2016 to finance the drilling into the active seismogenic fault zone of the Orkney earthquake and recover core samples of the faulted rock mass using the Moab Khotsong gold mine's underground workings as a platform. The purpose of the project, entitled "Drilling into Seismogenic Zones in Deep South Africa Gold Mines" (DSeis) was to investigate the Orkney earthquake active seismogenic fault zone to gain a better understanding of the geological structure on which the earthquake happened, why the earthquake happened, lithological, stress state in great depth, the earthquake mechanisms, seismic, and deep geomicrobiological analyses (Berset, 2017; Ogasawara et al., 2017; Ziegler et al., 2018). The project funding proposal was finally approved by the ICDP in August 2016 (Ogasawara et al., 2017).

The project team (DSeis) proposed to drill a borehole from the Moab Khotsong gold mine's underground to find and intersect the active fault zone of the Orkney earthquake, and then drill a fan of short holes from the end of the main borehole. A hydraulic drilling machine (LM90 from Boart Longyear Ltd) used for routine underground geological drilling was collared at a depth of 2.9 km on mining level 95 and targeted the upper fringe of the aftershock cloud. The borehole was drilled with an NQ size diamond-impregnated bit that allows cylindrical cores of solid rock to be retrieved (borehole diameter 75.7 mm; core diameter 47.6 mm). NQ is the minimum diameter that permits borehole geophysical logging. The borehole was drilled with a double barrel coring method and normal fluid circulation. The double barrel coring method was used to protect the cores from the circulating drilling fluid and from the drilling pipe rotation, almost complete core recovery was achieved (Berset, 2017; Ogasawara et al., 2019).

The drilling of the first borehole (borehole A) at an angle between 35^{o} and 45^{o} dipping northeast downwards of length 817 m from the borehole collar started in June 2017 and finished in September 2017, without intersecting the aftershock cloud of the Orkney active seismogenic zone (**Figure 1.3**). Borehole A did not intersect the Orkney active seismogenic fault zone due to the strong clockwise deflection from its proposed trajectory to run approximately 100 m and sub-parallel from the NNW-SSE striking aftershock cloud. However, borehole A did traverse the

periphery of the Orkney active seismogenic fault zone (Ziegler et al., 2018; Ogasawara et al., 2018; Ogasawara et al., 2019; **Figure 1.3**). Core samples were recovered continuously from borehole A. Downhole geophysical borehole logging (e.g., Gamma ray, density, borehole televiewer, P- and S- wave velocity, etc.) was carried out on borehole A in 2017 after the drilling had stopped (Rickenbacher, 2018).

The second borehole (borehole B) was collared at a depth of 2.9 km on mining level 95 similar to drilling location of borehole A. The drilling started in November 2017 and finished in June 2018 at a total length of about 700 m and angle of 45^{o} (**Figure 1.3**). At distance 612 m from the borehole collar, borehole B successfully intersected the upper edge of the aftershock cloud at a depth of about 3.3 km below surface, indicated by a core-loss zone of about 3.2 m wide (Ogasawara et al., 2018; 2019; **Figure 1.3**). Borehole B encountered core-loss because of the sharp lithological transition at borehole collar length of 612 m. The core-loss zone and aftershock cloud are spatially correlated, and this suggests that the zone may be the geological structure that hosted the 2014 Orkney earthquake (Ogasawara et al., 2018; 2019).

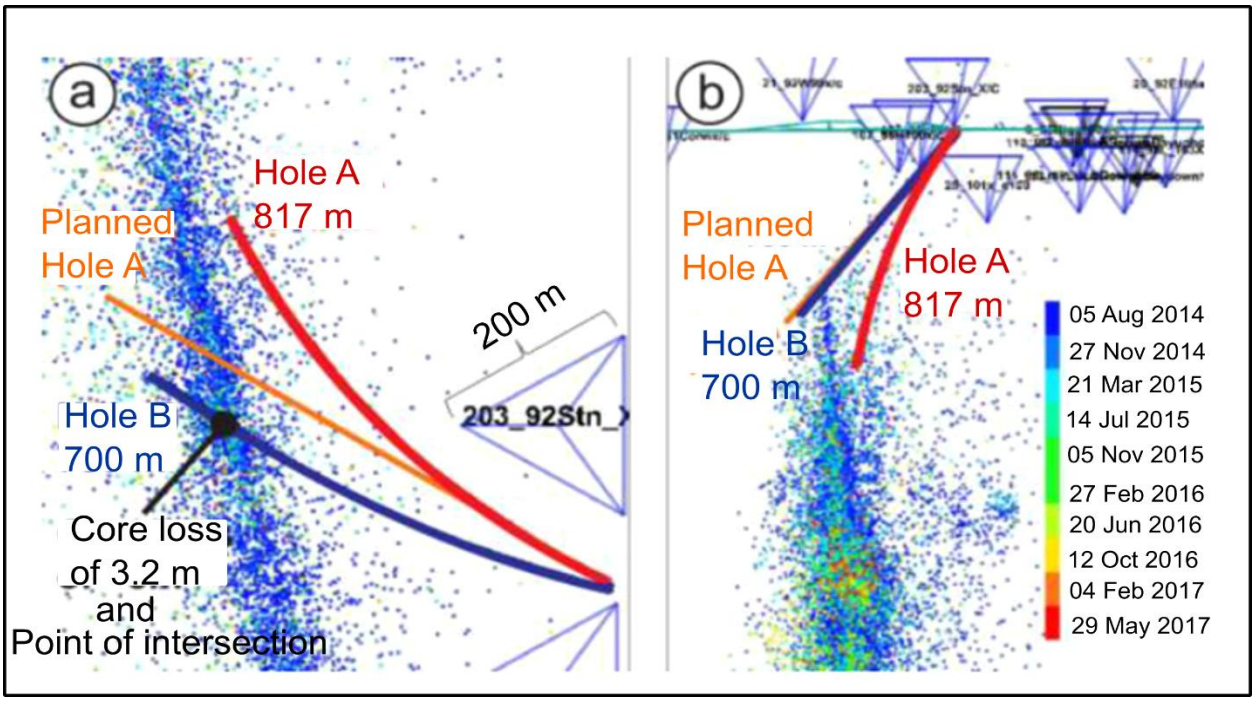

Figure 1.3: Views of the aftershock cloud, mining levels and DSeis boreholes. (a) Schematic top view. (b) Vertical section view. Borehole A drifted from the proposed trajectory without intersecting the aftershock cloud (from Ogasawara et al., 2018).

The double barrel coring tube system could not pass the unstable zone at length 612 m of borehole B to allow downhole geophysical borehole logging. In May 2018, a branch borehole (borehole C) was drilled ~ 544 m from the collar of borehole B at an angle of 2º-3º. Borehole C was drilled using the triple barrel coring tube system that enabled the broken rock fragments to be recovered and sampled and attained a total length of ~ 95 m (Rickenbacher, 2018; Ogasawara et al., 2018; 2019). Continuous recovery of core samples from borehole B and C was successful, with the exception of the core-loss zone where broken rock fragments were recovered.

The mine geologists were the first to visually describe the recovered core samples from boreholes A and B based on the rock type, mineralogy, stratigraphic horizon, grain size, hardness, and colour. After which a comprehensive core description was done by Berset, (2017); Ziegler et al. (2018); and Rickenbacher (2018) using thin sections. Berset, (2017) logged Hole A from the collar to a depth of about 213 m. All studies reveal that boreholes A and B intersected metasedimentary rocks (metasiltstone and quartzite) from the Upper and Lower Roodepoort Formation, quartzite strata

from the Babrosco Formation and flood basalts from the Crown lava Formation referred to as metabsalts, all members of the Jeppestown Subgroup of Archaean West Rand Group. Also, the studies reveal that all formations are intruded by diabase dykes and sills. Nkosi et al. (2022) performed laboratory physical measurements (bulk density, porosity, P- and S- waves velocities) on the core samples. Compared to the metabasalts and intrusive rocks, the metasediments exhibit the lowest bulk densities and seismic velocities. The Nkosi et al. (2022) study also noted slight overlaps in S-wave velocities and bulk densities between the metabasalts and metasediments. Nkosi et al. (2022) further found that laboratory estimated physical properties (bulk density, P- and S- waves velocities) conform with the downhole measurements in boreholes A and B.

Despite the availability of geological log data from the ICDP DSeis boreholes, no research has employed data-driven methods, such as machine learning-based algorithms for lithology classification in this context. To bridge this gap, this study demonstrates results of machine learning-based algorithms which are intended to offer insights into optimal model and identify key geophysical log features that contribute to accurate lithology classification. Ultimately, the findings aim to serve as a valuable control measure for future lithology classification studies based on geophysical log data.

1.4. Dissertation Structure

The rest of this dissertation is structured as follows:

- The regional and local geological background of our study is discussed in chapter **Error! Reference source not found.**.
- Geophysical logs and the source of the data are described in chapter 3.
- Our methodology is proposed and analyzed in chapter 4. Section 4.1 gives a general overview of machine learning and applied algorithms. Section 4.2 explains the methods used to achieve the classification of lithology in order to differentiate the different layers of quartzites and intrusives. Section 4.3 provides methods to determining the impact of correlation between logging on classification performance. Section 4.4 provides the

method to the determination of the optimal number of K for clustering of lithology in order to map the key radioactive lithologies.

- Results are presented in chapter 5. Section 5.1 and 5.2 interprets the results to the supervised classification of lithology. Section 5.3 provides results to the impact of the linear correlation between log features on the classification performance of the models in borehole A. Section 5.5 explains the results of K-means clustering in borehole A.

- Our results are explained in chapter 6. Section 6.1 discusses the results of the classification of lithology using supervised models. Section 6.2 discusses the results of the impact of the linear correlation between logging features on classification performance of a supervised models. Section 6.3 discusses the outcomes of K-means clustering for the mapping of key radioactive zones.

- Conclusions of the study are given in chapter 7.

2. REGIONAL AND LOCAL GEOLOGICAL BACKGROUND

2.1. The Witwatersrand Basin

The Witwatersrand Basin is an oval-shaped Archaean basin that is deposited on the central position of the Kaapvaal Craton in South Africa **(**Robb and Meyer, 1995; Frimmel et al., 2005). The Witwatersrand Basin is unique in terms of its mineral wealth and is of great importance to geoscientists and explorationists because it contains the world's oldest known well preserved and laterally extensive successions of sedimentary strata, as well as hosting the world's largest rich gold-bearing conglomerate beds (Armstrong et al., 1991; Frimmel and Minter, 2002; Frimmel et al., 2005). The Witwatersrand Basin hosts ten different goldfields as shown in Error! Reference source not found., namely: Free State, Kroonstad, Klerksdorp, West Wits Line, Central Rand, South Rand, Evander, East Rand, West Rand, and Potchefstroom Goldfields.

A change in the tectonic regime from compressional to transtensional on the Archaean Kaapvaal Craton has been proposed by Frimmel et al. (2005) to provide the depositional environment for the Witwatersrand Basin. The development stages of the Witwatersrand Basin can be differentiated based on the available three stratigraphic sequences which follow a traditional stratigraphic progression (Armstrong et al., 1991; Frimmel and Gartz, 1997). The stratigraphic succession of the Witwatersrand Basin is called the Witwatersrand Supergroup. The Witwatersrand Supergroup's first development stage entailed crustal extension followed by the unconformable deposition of the Dominion Group (3.07–3.09 Ga) on an Archaean granite-greenstone basement making up the Kaapvaal Craton (Armstrong et al., 1991; Frimmel and Gartz, 1997). The second stage involved epicontinental-style sedimentation of the West Rand Group (2.9–2.97 Ga). During the last stage, compression and uplift of the hinterland produced fluvial sediments, including gold-bearing conglomerate beds, of the Central Rand Group (2.71–2.89 Ga), (Armstrong et al., 1991; Robb and Meyer, 1995; Frimmel et al., 2005).

The lowermost group of the Witwatersrand Supergroup is the West Rand Group (Error! Reference source not found.). The West Rand Group thins to the northeast and has an average thickness of about 5 km in the Klerksdorp region (Frimmel et al., 2005), (Error! Reference source not found.). The West Rand Group comprises cycles of metamorphosed marine shales, conglomerate beds,

shallow marine quartzites, and minor banded iron formations (Armstrong et al., 1991; Els, 2000), ( ). The West Rand Group is subdivided into three units (the Jeppestown, Government, and Hospital Hill subgroups) based on the shale/sandstone ratio variation and basin-wide disconformities (Frimmel et al., 2005). The Jeppestown and Government subgroups include minor gold-bearing conglomerate beds (Frimmel et al., 2005), .

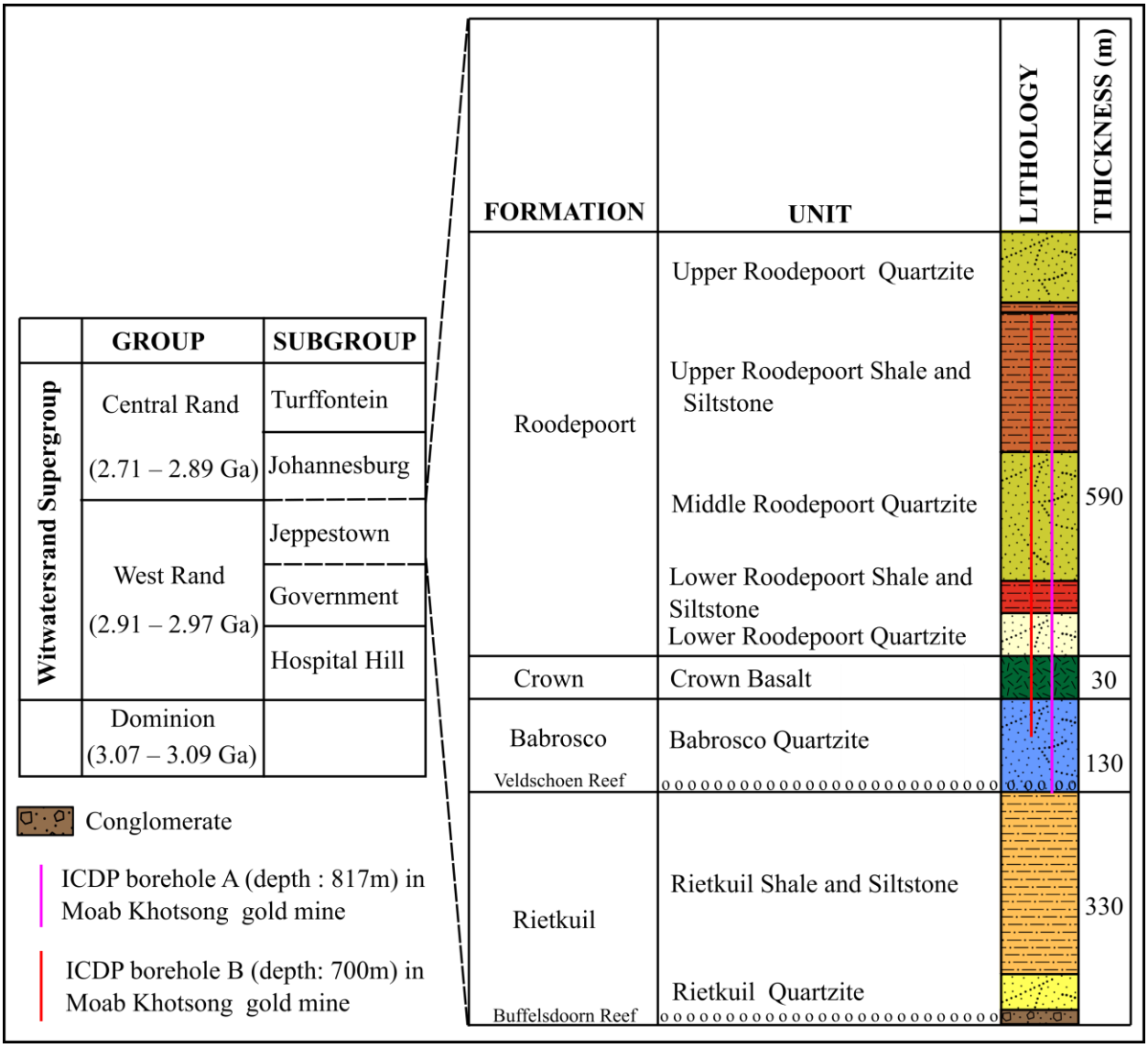

Figure 2.1: Simplified stratigraphic column of the Jeppestown Subgroup drilled by the ICDP DSeis boreholes at the Moab Khotsong gold mine (modified after Robb and Meyer, 1995).

The West Rand Group minimum age (2.91 Ga) is given by the amygdaloidal basaltic unit in the Crown Formation, the only recognized volcanic unit of about 250 m thick within the Jeppestown Subgroup (Armstrong et al., 1991; Frimmel et al., 2005). The maximum age (2.99 Ga) of the West Rand Group was provided by the analysis of Uranium-Pb data from detrital zircon grains (Robb and Meyer, 1995).

The Central Rand Group is the uppermost group of the Witwatersrand Supergroup. The Central Rand Group overlies the West Rand Group unconformably and attains a maximum thickness of about 3 km near the center of the basin (Frimmel et al., 2005), (Error! Reference source not found.). The Central Rand Group is characterized by quartzites, occasional shales, and gold-bearing conglomerate beds such as the Vaal- and C- Reef, which were deposited in a fluvial and shallow marine environment (Law et al., 1990; Armstrong et al., 1991; Frimmel and Minter, 2002; Robb and Meyer, 1995). The remarkable richness in gold concentration of the Witwatersrand Basin is confined to the Central Rand Group (Frimmel et al., 2005). There are two subgroups of the Central Rand Group: The Turffontein and Johannesburg groups, both predominately quartzite. Volcanic activities in the Central Rand Group are indicated by the minor (about 170 m thick) amygdaloidal basalt units (Bird Member) in the Krugersdorp Formation within the Johannesburg Subgroup (Armstrong et al., 1991; Frimmel and Minter, 2002; Nkosi et al., 2022). The structure of the Witwatersrand Basin is complex. A series of tectonic events produced various unconformities, faults, sills, dykes, and mafic intrusion (Robb and Mayer, 1995; Catuneanu and Biddulph, 2001; Dankert and Hein, 2010).

2.2. Geological overview of The ICDP DSeis Boreholes

The ICDP DSeis project at Moab Khotsong gold mine drilled through the West Rand Group's Jeppestown Subgroup ( ). Boreholes A and B intersected the metasiltstone and quartzitic strata from the Roodepoort and Babrosco formations, which are separated by a conformable igneous rock layer called the Crown Formation. In addition, the Roodepoort, Babrosco, and Crown formations are intruded by intrusives (Error! Reference source not found.) of unknown age, varying in thickness and colors, consisting of sills that run sub-parallel to the bedding, and dykes that dip at high angles (Armstrong et al., 1991; Meier et al., 2009; Ziegler et al., 2018; Rickenbacher, 2018).

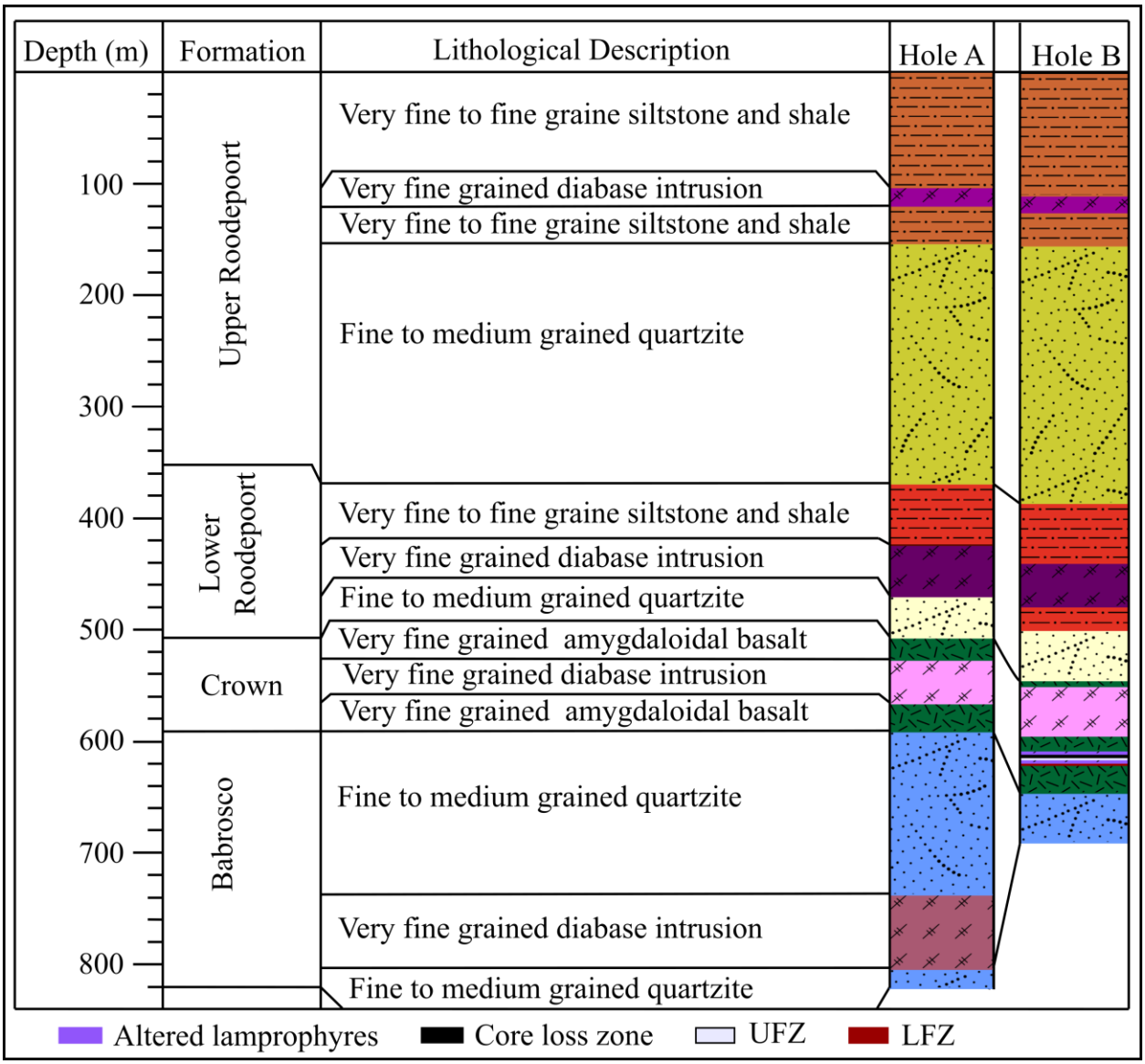

Figure 2.2: The geological log overview of the ICDP DSeis boreholes at Moab Khotsong gold mine. Profile sections are colour-coded based on lithology. UFZ and LFZ denotes the upper fault zone and lower fault zone, respectively.

The Roodepoort Formation extends from the DSeis boreholes collar depth at 2.9 km below surface to a drilled borehole length of 504 m and 543 m in Hole A and Hole B, respectively (Ziegler et al., 2018; Rickenbacher, 2018). The minimum true thickness is computed to be 483 m based on the borehole orientation and dip of the strata. The total thickness of the Roodepoort Formation is bigger since drilling started in the Upper Roodepoort Formation. The Roodepoort Formation is subdivided into Upper and Lower layers. Both the Upper and Lower sections of the Roodepoort

Formation consist mainly of alternating laminated meta-sedimentary rock types: metasiltstones with interbedded metamorphosed shale, and quartzite (Mambane et al., 2011; Ziegler et al., 2018; Rickenbacher, 2018). Ziegler et al. (2018) and Rickenbacher (2018) reports about 95 m and 55 m thickness for Upper and Lower Roodepoort formations, respectively

According to the thin sections of Berset (2017) and Rickenbacher (2018), the metasiltstone from both the Upper and Lower Roodepoort formation in Holes A and B is grey to dark grey coloured, very fine to fine grained metasiltstone with frequent interbedding of dark green bands of shale **(**Error! Reference source not found.). About 80% of the metasiltstone volume comprises medium grains of quartz (proto- and ortho-quartzite) and plagioclase of about 15-60 µm diameter. Approximately 15% of the volume is occupied by minor mineral components such as sulphides, which create secondary crystallized pyrite and chalcopyrite. Sulphides appear often along the quartz veins. The rest of the volume (5%) is occupied by amphibole. Generally, the metasiltstone appears very heterogeneous. The strong heterogeneity can be explained by the frequent bands of shale in metasiltstone.

Similarly, two units of fine to medium grained, light grey or greenish grey cross-bedded quartzites of about 205 m and 35 m thick were intersected in the Upper and Lower Roodepoort formations, respectively, in Holes A and B (Error! Reference source not found.). The quartzites intersected in the Roodepoort Formation are stratified and cross-stratified. The thin sections studied by Rickenbacher (2018) show that quartz grains (orthoquartzite and protoquartzite) are the major mineral component occupying more than 90% of the total volume of quartzite in these formations; while plagioclase feldspar, calcite, and euhedral secondary crystallized sulphides are the minor mineral components occupying the remaining 10% (Ziegler et al., 2018; Rickenbacher, 2018).

At ICDP DSeis boreholes, Upper and Lower Roodepoort formations starts with the metasiltstone lithology and ends with the quartzite lithology as seen in Error! Reference source not found.. Therefore, the transitional lithological boundary between the metasiltstones and quartzites can be used to distinguish the Upper Roodepoort Formation layer from the Lower Roodepoort Formation layer (Rickenbacher, 2018). Changes in grain sizes from very fine (metasiltstone) to medium grained (quartzite), and also changes in colour from grey, dark grey (metasiltstone) to light or

greenish grey (quartzite) may also be indicators of the lithological boundary between the Upper and Lower Roodepoort formations.

The rock strata from the Crown Formation intersected by Holes A and B are amygdaloidal basalts dated at 2.91 Ga using the Zircon U-Pb dating technique (Armstrong et al., 1991; Mambane et al., 2011). The Crown Formation has between 87 m and 101 m apparent thickness along Hole A and Hole B, respectively. The true thickness of Crown Formation is computed to be about 80 m. The amygdaloidal basalts are mafic rocks with homogenous matrix which are olive green to dark grey in colour and very fine grained. The thin section study of Rickenbacher (2018) shows that amygdaloidal basalt has plagioclase and pyroxene grains as the major minerals. Accessory minerals include magnetite**,** quartz grains, calcite, pyrite, and chalcopyrite.

The Babrosco Formation is the deepest formation intersected by Holes A and B. The Babrosco Formation exhibits an apparent thickness of 226 m and 55 m along Holes A and B, respectively. The true thickness of Babrosco Formation is computed to be about 198 m. The rock strata from the Babrosco Formation are light grey, fine to medium grained feldspathic quartzite (Armstrong et al., 1991; Ziegler et al., 2018; Rickenbacher, 2018; Nkosi et al., 2022). The quartzites of the Babrosco Formation are similar to the quartzites of the Upper and Lower Roodepoort formations in mineral composition, texture, color, and structure according to the thin sections results of Rickenbacher (2018).

The thin sections studied by Ziegler et al. (2018) and Rickenbacher (2018) on the diamond core drill core samples recovered from Holes A and B reveals intrusive rocks that are very fine-grained close to the contact with the host rock, but medium grained in the middle. The very fine-grained rock close to contact with the host rock is probably because of the fast cooling at that point. The intrusive rocks were originally fine to medium grained, eqigranular basic rocks of gabbroic composition, but due to the strong alteration by fluids, the intrusive rocks are classified as diabase. A diabase is a basic, holocrystalline, and subvolcanic rock like an altered gabbro. Plagioclase and pyroxene are the major mineral components of the intrusive rocks. The minor mineral components were revealed to be calcite, quartz, apatite, and epidote with the presence of accessory minerals like pyrite, chalcopyrite, magnetite, and titanium oxide. Rickenbacher's (2018) thin sections

further reveal that the major mineral components (plagioclase and pyroxene) of the intrusive rocks were strongly altered by fluids and recrystallization. In some cases, the major minerals were totally replaced by secondary minerals such as amphibole, chlorine, and actinolite.

The intrusive rocks consist of similar mineral composition. However, only the grade of alteration varies over the different intrusive rocks across the different formations. In addition, the intrusive rock at the core-loss zone within the Crown formation (610.94-613.10m) in Hole B is optically different from the other intrusives. This difference is because of the change in colour from the normal greenish grey colour to black colour, which arises as one of the main mineral components (pyroxene) is replaced by very fine-grained sericite minerals (higher content of mica). Hence, the light is absorbed by the replaced sericite minerals, which caused this intrusive at the core-loss zone to be dark grey to black in colour (Ziegler et al., 2018; Rickenbacher, 2018). Miyamoto et al. (2022) noted that the recovered broken rock fragments close to the core-loss zone and the fault zones located within the Crown Formation (610.94-619.94m; **Figure 2.2**) display mineral composition (high amphibole, biotite, talc, calcite contents with high potassium mafic) of an altered lamprophyre intrusions. In addition, Miyamoto et al. (2022) suggested that the altered lamprophyres may have hosted the 2014 Orkney earthquake, but frictional complexity may have controlled the magnitudes of the mainshock, and the distribution of the aftershocks.

In summary, a total of ten classes of lithologies (**It** is important to accurately classify these lithologies between layers at IDCP DSeis boreholes situated at the Jeppestown subgroup of the West Rand Group of Witwatersrand Supergroup for safety mapping and objective decision making.

Table 2.1): Upper Roodepoort quartzite, Upper Roodepoort siltstone, Upper Roodepoort diabase intrusion, Lower Roodepoort siltstone, Lower Roodepoort quartzite, Lower Roodepoort diabase intrusion, Crown basalt, Crown diabase intrusion, Babrosco quartzite, and Babrosco diabase intrusion based on their stratigraphic positions, which were recorded from the ICDP DSeis

boreholes using traditional methods (thin sections and core observation), (Ziegler et al., 2018; Rickenbacher, 2018) are the expected to be predicted by machine learning models. It is important to accurately classify these lithologies between layers at IDCP DSeis boreholes situated at the Jeppestown subgroup of the West Rand Group of Witwatersrand Supergroup for safety mapping and objective decision making.

Table 2.1: Classes of lithologies and data size from DSeis boreholes A and B.

Borehole A		Borehole B	
Lithology	Data size	Lithology	Data size
Upper Roodepoort quartzite	4147	Upper Roodepoort quartzite	4721
Lower Roodepoort siltstone	1107	Upper Roodepoort siltstone	1012
Lower Roodepoort diabase intrusion	942	Upper Roodepoort diabase intrusion	300
Lower Roodepoort quartzite	711	Lower Roodepoort siltstone	1059
Crown basalt	898	Lower Roodepoort diabase intrusion	741
Crown diabase intrusion	783	Lower Roodepoort quartzite	1300
Babrosco quartzite	3183	Crown basalt	200
Babrosco diabase intrusion	1323	Crown diabase intrusion	1040

3. GEOPHYSICAL LOGGING AND DATASET USED

Geophysical logs are digital records of continuous variations in physical properties of a rock mass with depth in borehole(s) drilled for the purpose of hydrocarbon, geothermal, minerals or water exploration, measured using downhole geophysical sensors (sonde). Grain size, mineral composition, porosity, presence of pore fluids, effective pressure, cementation and compaction, among other parameters all have a significant impact on the rock physical properties (Mostaghel, 1999; Huang and Pan, 2004; Bougher, 2016). Hence, within borehole(s), variation in geophysical log recordings denotes a change in physical properties, which depends on the targeted subsurface lithology.

3.1. Acquisition

The act of acquiring geophysical logs data is called geophysical logging. In geophysical logging, a sonde (e.g., **Figure 3.1**) is moved in or out of a borehole, and measurements are recorded at a regular interval. The digitally recorded data from the subsurface sonde is transmitted through a cable to the recorder at the collar of the borehole, where it is stored and analyzed (Mostaghel, 1999).

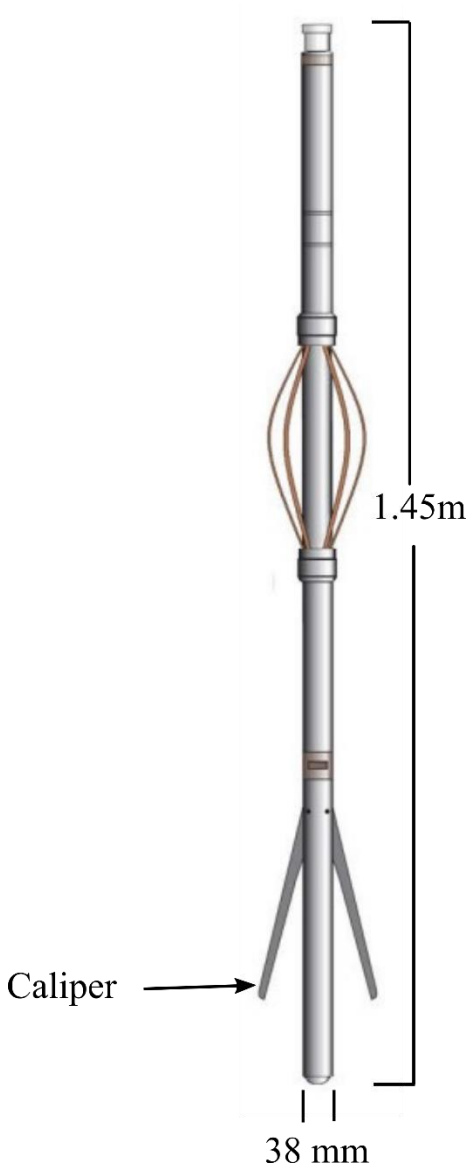

Figure 3.1: A 3-arm caliper logging sonde (from Chatfield, 2017).

Marcel and Conrad Schlumberger conducted the first commercial geophysical well logging in 1927 while studying the principle of "electrical coring" in an oilfield at Merkwiller-Pechelbronn, in eastern France. Marcel and Conrad conducted the logging with a resistivity tool to identify hydrocarbon-bearing layers using difference in the porosity of sandstone lithology. From this study, they concluded that the technique was reliable for identifying hydrocarbon-bearing layers due to the resistivity contrast between sandstone layers. They further postulated that the technique could be more reliable for geological correlation purposes than drilling cuttings, and less expensive than core drilling (Luthi, 2001; Chopra et al., 2002).

Geophysical logging is extensively used in the mining sector for mineral exploration activities, assessment of ore reserves, mining associated risk mitigation, and detection of economical mineral deposits (e.g., uranium ores). Furthermore, groundwater exploration and evaluation use geophysical logging to define aquifers and producing zones (Luthi, 2001). It is often more reliable to classify the lithological units surrounding a borehole using geophysical log data than core logging or drilling cuttings. This is because geophysical logging offers reliable continuous and numerous *in-situ* rock physical properties that can be integrated to characterize the rock mass (Huang and Pan, 2004). Also, geophysical logging offers an opportunity to sample a larger volume of the borehole surroundings than drilled core samples (Mostaghel, 1999; Luthi, 2001).

However, reliable classification of lithology using geophysical logs data could sometimes be challenging, non-linear and complex. These may be due to the influences of drilling fluids and additives on the walls of the borehole, grain size, the rock types that got altered and the area the rock fragments (sediments) passed through before deposition, pore fluids, pore shape, fluid saturation, inconsistency in acquiring geophysical logs, and effective pressure etc. (Gueguen and Palciauskas, 1994; Luthi, 2001; Krygowski, 2003; Alves et al., 2014; Tiab and Donaldson, 2015; Hayat et al., 2019).

3.2. Applied Geophysical Logs

This section of the study briefly explained the geophysical logs applied in this study (natural gamma ray, density, $V_{p,}$ and V_s). Caliper sonde is the first logging technique run through a borehole to indicate the borehole conditions and stability before sondes that are more costly, fragile or containing sources of radioactivity are run through the borehole. This is because data reliability is degraded by bad borehole conditions (Stefansson and Steingrimsson, 1990; Mostaghel, 1999). However, caliper log data was not provided for this study. A summary of the geophysical logging at ICDP DSeis boreholes is listed in section **3.3.**

3.2.1. Total Gamma Ray Log

The total gamma ray log provides measurement of continuous variation of total natural occurring gamma radiation emitted by radioelements from rock bodies surrounding borehole(s) with the

depth. The emitted gamma ray progressively reduces energy as a result of collisions with other atoms in the rock formation (i.e., Compton scattering); (Chatterjee and Paul, 2012; Fiser-Nagy et al., 2014; Nazeer et al., 2016; Glover, 2000). Compton scattering continues until the energy of the emitted gamma ray is low enough for the rock formation to absorb completely. Hence, the measured gamma ray intensity is a function of (Schlumberger, 1989; Mostaghel, 1999; Chatterjee and Paul, 2012; Fiser-Nagy et al., 2014; Nazeer et al., 2016; Glover, 2000):

- The initial intensity of gamma ray emission, and
- The amount of Compton scattering encountered by the gamma rays, which is determined by the detector, the intervening material density, and the distance between the gamma emission.

The gamma ray logging sonde consists of a highly sensitive gamma ray detector in the form of a scintillation counter. The scintillation counter composed of a single sodium iodide crystal activated by thalium, which is supported by a photomultiplier that is made of a photocathode and a series of anodes held at increasingly higher electrical potentials, all serially arranged in a high vacuum. When the sodium iodide is struck by a gamma ray, a little flash of light is produced. Since the flash is too weak to be detected by traditional electronics, the photomultiplier amplifies it. When the flash of light strikes the photocathode of the photomultiplier, many primary electrons are created. The primary electrons accelerate in the initial anode's direction. Four to eight secondary electrons are released for every primary electron that strikes the anode. The next anode's direction is where these released electrons are accelerated, and each secondary electron there produces more secondary electrons (Glover, 2000). For each incident electron, if 6 electrons are emitted at each anode, one incident gamma ray can be said to produce 6^{10} = 60,466,176 electrons, which creates a current that can be further amplified using traditional amplifiers (Glover, 2000). The photomultiplier is saturated and unresponsive to additional gamma rays during this short process period. However, saturation is rarely seen in downhole logging due to the small incident flux of gamma rays generated from rocks. The energy of the incident gamma ray is proportional to the final current from the scintillation counter since the gamma ray's energy is proportional to both the number of primary electrons and the light flash. Hence, below a certain gamma ray energy, the

scintillation counter will become insensitive to gamma rays. The gamma ray logging tool has a high vertical resolution, since the scintillation counters are relatively small devices (Glover, 2000). The depth penetration of gamma ray logging tool depends on the density of the formation and drilling mud. For example, for average density values of formation and drilling mud, the depth of investigation is 30 cm if about 75% of gamma ray signal is defined. According to Glover (2000), the depth of penetration will be lower for denser drilling mud and formations, and the depth of penetration will be higher for less dense drilling mud and formations.

The calibration of the gamma log can be in API units (API = American Petroleum Institute), or gamma counts per second. The three elements: Potassium (40K), Uranium (238U), Thorium (232Th) and their daughter products in their decay are the common sources of radiation found throughout the Earth's crust (Luthi, 2001). Gamma ray logs are known to be effectively applicable in mapping lithologies that contain radioactive elements. Note that shales, clay and potassium feldspar minerals will mostly show high gamma ray values in rock materials surrounding borehole(s) (Fiser-Nagy et al., 2014; Nazeer et al., 2016). Therefore, metasiltstone with interbedded shale and quartzite that contains feldspar may be easily identified by their higher gamma ray values. The gamma ray logs are also used to indicate certain radioactive mineral deposits with high natural radioactivity (e.g., potash and uranium ores) and non-radioactive mineral deposits with low natural radioactivity (e.g., coal beds), (Luthi, 2001; Huang and Pan, 2004; Glover, 2000).

The application of natural gamma ray tool has a significant advantage of not being affected by borehole conditions and may be run through casing or drill pipe (Chatterjee and Paul, 2012; Fiser-Nagy et al., 2014; Nazeer et al., 2016). However, the gamma ray tool will show a significant decrease when run through a casing. Consequently, gamma ray logs can reliably be used to match the depth of data from a given depth interval made at different times with different logging tool. relying on the characteristic sudden decrease in gamma ray values when the tool comes into contact with the borehole casing, section above the interval of interest. But depth matching mostly relies on matching the patterns in the gamma ray signature from the gamma ray tools run with each combined tool (Glover, 2000).

3.2.2. Bulk Density Log

Density log provides a continuous record of the bulk density of rock formations surrounding the borehole. Density logging is related to the Compton scattering phenomenon, where gamma rays emitted by a radioactive source within the sonde are scattered in all directions by colliding with several electrons in the formation's atoms. Compton scattering gradually reduces the energy of the gamma rays. When the energy of the emitted gamma rays is less than 0.5 MeV, they are ultimately absorbed by the formation (Glover, 2000). The density logging tools consists of a radioactive source that bombards the rock formation with radiation and measures the number of scattered radiations by the formation returning to the logging tool along the depth of borehole(s). The tool also consists of a short-space detector placed 7 inches from the radioactive source that acts similarly to the detectors used in the total natural gamma ray tools, and a long-space detector placed 16 inches from the radioactive source to compensate for challenges associated with the mudcake (Luthi, 2001; Huang and Pan, 2004; Glover, 2000).

The density logging tool operation depends on the ability to identify which gamma rays have been scattered after being supplied to the formation by the tool's radiation source. The short-spacing detector identifies gamma rays that have only penetrated a short distance into the formation and mudcake before scattering back to the detector. Therefore, the short-spacing detector readings are measures of attenuation in the near-borehole region (i.e., mudcake and very shallow region in the formation). While the long-spacing detector identifies gamma rays that have penetrated through the mudcake and deeper into the formation before scattering back to the detector. Thus, the long-spacing readings are measures of attenuation in the formation (Glover, 2000). The calibration scale of the bulk density log can be in g/cm^3 units. The density logging tool is sensitive to borehole quality since it investigates shallow depth (Glover, 2000). Needless contact with the radioactive source at the surface should be avoided.

Common applications of density logs are (Luthi, 2001; Huang and Pan, 2004):

- Estimation of rock mechanical properties and overburden pressure,
- Identification of lithologies when used together with other logs (especially with neutron log),

- Indicator of gas-bearing zones,
- Estimation of reflection coefficients and acoustic impedance to generate synthetic seismic profiles and to assess the seismic horizons depths, when used together with acoustic velocity from sonic log,
- Identification of minerals, and
- Estimation of reservoir parameters such as porosity and permeability.

3.2.3. Sonic log

Sonic logs provide a continuous record of variation of travel time of acoustic waves (i.e., P- and S- waves) through rocks surrounding borehole(s) wall with depth. Sonic log can be used to identify lithology since travel time of acoustic waves through rocks varies notably with lithology and rock texture. The sonic log can be applied to estimate porosity of formations which is an essential parameter for the exploration of hydrocarbons. In a hard rock mining environment, sonic logs are vital as they provide 'seismic' velocity date use for the creation of synthetic seismograms and use with density log to calculate the reflectivity profile for future seismic study. Sonic logs are also used to estimate the rigidity, pore fluid content, degree of fracture, and confining stress of rocks surrounding borehole(s), (Luthi, 2001; Huang and Pan, 2004).

The main components of a sonic logging tool are mostly one transmitter and two receivers: one located close to the transmitter and one far away. The two receivers were made to take into consideration the time the elastic waves take to travel through the drilling mud, which is included in the measurement of the travel time and caused the measured travel time to be too long. To achieve this, the tool's transmitter and receivers are spaced apart, usually between 3 and 5 feet. Also, the two receivers were designed to adjust for the altered critical refraction angle caused by the waves' velocity variations because the elastic wave traveled through a non-constant length of formation. In order for the P-wave from the virgin formation to arrive before the one from the altered zone in an altered zone with low velocity, there needs to be large distance (about 3 to 5 feet) between the transmitter and the receiver.) A Long Spaced Sonic (LSS) tool provides better data than a Borehole Compensated Sonic (BHC) tool in an altered zone with low velocity. While for an altered zone with high velocity, any kind of sonic logging tool can be used, and the measured

log data will be that of an altered zone (Glover, 2000). The two receivers are used to measure the difference in time the elastic waves take to arrive at each receiver from a given pulse from the transmitter. This time is known as the sonic interval transit time (Glover, 2000; Luthi, 2001; Huang and Pan, 2004). The transmitter and the receivers are separated by a low velocity and high attenuation material such as a rubber connector that can reduce the amount of direct transmission of acoustic waves along the sonde from transmitter to receiver" (Luthi, 2001).

The depth penetration of sonic logging tool depends on the wavelength of the elastic wave and not on the spacing between the transmitter and the receivers. For any given sonic logging tool, the higher the formation velocity, the larger the wavelength, and the deeper the penetration (Glover, 2000).

3.3. ICDP DSeis Boreholes Geophysical Logging at Moab Khotsong Gold Mine

The first geophysical logging was conducted in Hole A from 0-180 m collar depth on 28th and 29th June 2017 (Berset, 2017). Hole A of length 800 m was completely logged in October 2017, and Hole B was logged in November 2018 (Rickenbacher, 2018). The two boreholes were cleared off of all drilling mud by flushing the boreholes with fresh water and filling them completely with clear water to prepare them for downhole logging (Berset, 2017). The downhole sondes were run at 6 m per minute for the fastest except for the OPTV sonde that was ran between 1.5 – 2 m per minute to have a high vertical resolution of wall images. The sondes were run one after the other to acquire non-disturbed data. The logging company, Digital Surveying provided a specific winch and steel cable to ensure the logging speed is controlled (Berset, 2017). Specification of tools used for the ICDP DSeis downhole geophysical logging at Moab Khotsong gold mine conducted by Digital Surveying is summarized in **Table 3.1**.

The Full Wave sonic sonde (FWFS60) employed a dual transmitter and dual receiver array to provide high quality V_p and V_s data. The sidewall density sonde, which is pressed against the sidewalls of the borehole contains a radioactive source of Caesium 137, an active source of gamma ray. The sidewall density sonde is used to provide bulk densities of the different lithologies (Berset, 2017).

Table 3.1: Specification of tools used for the downhole geophysical logging of ICDP DSeis boreholes conducted by Digital Surveying.

Logging sonde	Data measured	Sampling interval (m)	Utilization
Natural Gamma	Natural γ-radiation	0.01	Depth matching, lithology identification
Sidewall density	Short spaced bulk density	0.01	Calculation of vertical stress
Full Wave Triple Sonic (FWFS60 slimhole)	P- wave and S-wave velocities	0.05	Estimate dynamic elastic properties

When the sonic log was first inspected, it showed that the S-waves signal was more amplified than that of P-waves. Therefore, every single-shot recorded was processed for S-wave velocity. The P-wave data was further subjected to a low-pass filter to improve the signal and distinguish it from the noise preceding P-wave (Nkosi et al., 2022). In Hole A, preceding the onset of P-wave, the rough texture of the hole sidewall caused noise resulting to challenges in processing the P-wave velocity data from 0-160 m from the collar. To mitigate the noise impact, shots were recorded with receivers placed at 600 mm , 900 mm, and 1200 mm distances. P-wave velocity data was processed by stacking 10 recorded shots. In Hole B, P-wave velocity data was not processed due to the significant poor signal of P-waves (Rickenbacher, 2018; Ogasawara et al., 2019; Nkosi et al., 2022).

3.3.1. Dataset

The downhole logging of DSeis boreholes at Moab Khotsong gold mine provided opportunity to assess the automated classification of lithology using data-driven methods such as machine learning-based algorithms based on geophysical lo data with less reliance on the experience and knowledge of geoscientists. Geophysical log data offers advantages such as high spatial resolution, reliable continuity of *in-situ* data that are significantly large for lithology identification, and acquisition convenience compared to traditional lithology identification methods such as core analysis (Huang and Pan, 2004; Xie et al., 2018). Also, a downhole geophysical log dataset

identifies parts of the borehole with no core recovery (Benaouda et al., 1999). However, geophysical log data can be complex, non-linear and with errors. These may be due to the influences of drilling fluids and additives on the walls of the borehole, grain size, altered rock types and the environment the rock fragments (sediments) passed through before deposition, pore fluids, pore shape, fluid saturation, inconsistency in acquiring geophysical logs, and effective pressure etc. (Gueguen and Palciauskas, 1994; Luthi, 2001; Krygowski, 2003; Alves et al., 2014; Tiab and Donaldson, 2015; Hayat et al., 2019). Hence, they are carefully preprocessed before their application in lithology identification, especially when using machine learning algorithms. The dataset utilized in this study, available in Excel format, consists of downhole geophysical log (**Table 3.2** and

Table 3.3 and lithological logs (**It** is important to accurately classify these lithologies between layers at IDCP DSeis boreholes situated at the Jeppestown subgroup of the West Rand Group of Witwatersrand Supergroup for safety mapping and objective decision making.

Table 2.1) obtained using traditional methods and core samples from drilled ICDP DSeis boreholes in the Moab Khotsong gold mine. The dataset was processed and provided by the ICDP DSeis team members in Japan.

Table 3.2: Overview of applied geophysical log data from Hole A.

Sampled depth from collar (m)	Log data	Sampling interval (m)	Data size	Missing value within recorded values
1-817.24	Natural total gamma ray	0.01	81625	0
5.01-816.21	Short-spaced bulk density	0.01	81121	400
8.2-816.22	S-wave velocity (V_s)	0.05	19203	800
161.01-815.77	P-wave velocity (V_p)	0.05	13091	100

Table 3.3: Overview of recorded geophysical log data from Hole B.

Sampled depth from collar (m)	Log data	Sampling interval (m)	Data size	Missing value within recorded values
1-610.04	Natural total gamma ray	0.01	60905	0
89.39-607.99	Short-spaced density	0.01	51861	500
49.02-610.13	S-wave velocity (V_s)	0.05	11247	800

The combination of bulk density, Vp and Vs log data can best be applied to identify the various layers of the diabase intrusions and Crown basalt due to their distinct high values. The layers of the quartzites and siltstone may be easily identified using the gamma ray log due to the presence of clay minerals and feldspar in these lithologies, which causes a high response to gamma ray signature. The statistical distribution of the available log features data for each lithology type in borehole A is demonstrated in **Table 3.4Error! Reference source not found.**. More tables for other lithology type are shown in **Appendix A** and **Appendix B**.

Table 3.4: Summary overview of the statistical distribution of the log features for each lithology type in borehole A. 25%, 50%, and 75% denotes the first quartile, second quartile/median, and third quartile of logging data for each lithological class. Min and Max denoted the minimum and maximum value of logging data for each lithological class.

	Upper Roodepoort quartzite				Lower Roodepoort siltstone				Lower Roodepoort diabase intrusion			
	V_p (m/s)	GR (API)	V_s (m/s)	RHOB (g/cm^3)	V_p (m/s)	GR (API)	V_s (m/s)	RHOB (g/cm^3)	V_p (m/s)	GR (API)	V_s (m/s)	RHOB (g/cm^3)
Data size	4147	4147	4147	4147	1107	1107	1107	1107	942	942	942	942
Mean	5674	76.86	3698.49	2.68	5901	67.26	3618.42	2.79	6507	220.40	3883.60	3.02
Std	127.4	40.00	58.04	0.04	47.78	11.80	55.76	0.07	201.64	182.46	87.12	0.12
Min	5355	21.15	3537.74	2.58	5769	38.51	3473.03	2.66	6295	37.08	3676.47	2.87
25%	5589	58.49	3658.54	2.66	5867	59.00	3579.95	2.74	6494	90.87	3841.11	3.00
50%	5660	69.79	3699.59	2.68	5904	65.90	3617.95	2.77	6578	156.30	3910.83	3.06

75%	5750	84.66	3739.49	2.71	5933	74.06	3653.64	2.82	6627	257.50	3952.57	3.09
Max	5989	123.74	3856.04	2.78	6023	96.62	3764.12	2.92	6716	492.22	4000	3.23

High gamma ray values recorded at a distance of about 420 - 460 m from the collar of borehole A, denotes a section with a relatively high content of radioactive elements (marked by blue lines in **Figure 3.2a**). Studies by Yabe et al. (2019) reported that this section is saturated with high salinity water seeping from a fracture. **Figure 3.2a** demonstrates a strong correlation between the density log and the V_p log. The measured gamma ray value is very low (≤ 0 API) in the intruded diabase. The high gamma ray values (ranging between 50- and 600- API) measured in the classes of quartzite with a significant variability might be associated with the feldspar mineral component in quartzite. The gamma ray values of metasiltstone are medium (about 50 API) with less variability. The medium value of gamma ray may be due to the interbedded shale. Furthermore, **Figure 3.2** show that the intruded diabase has the highest density value, almost reaching 3.2 g/cm^3 and steady. The density of metasiltstones and quartzites is similar, with values ranging between 2.6- and 2.8-g/cm^3. The Crown basalt shows a much steadier density profile ranging between 2.7- and 2.8-g/cm^3. The intruded diabase lithology also has a smooth V_p and V_s profiles, with a highest value of 6700 m/s for V_p and 4000 m/s for V_s. The quartzite classes displayed the lowest value of V_p (between 5500 m/s and 5800 m/s) and V_s (between 3600 m/s and 3900 m/s) with less variability. The measured V_p and V_s in metasiltstone is similar to the quartzites with less variability and higher values between 5800 m/s and 6000 m/s for V_p and 3500 m/s and 3750 m/s for V_s. The Crown basalt has smoother and higher V_p and V_s profiles, with values of 6100 m/s and 3600 m/s, respectively.

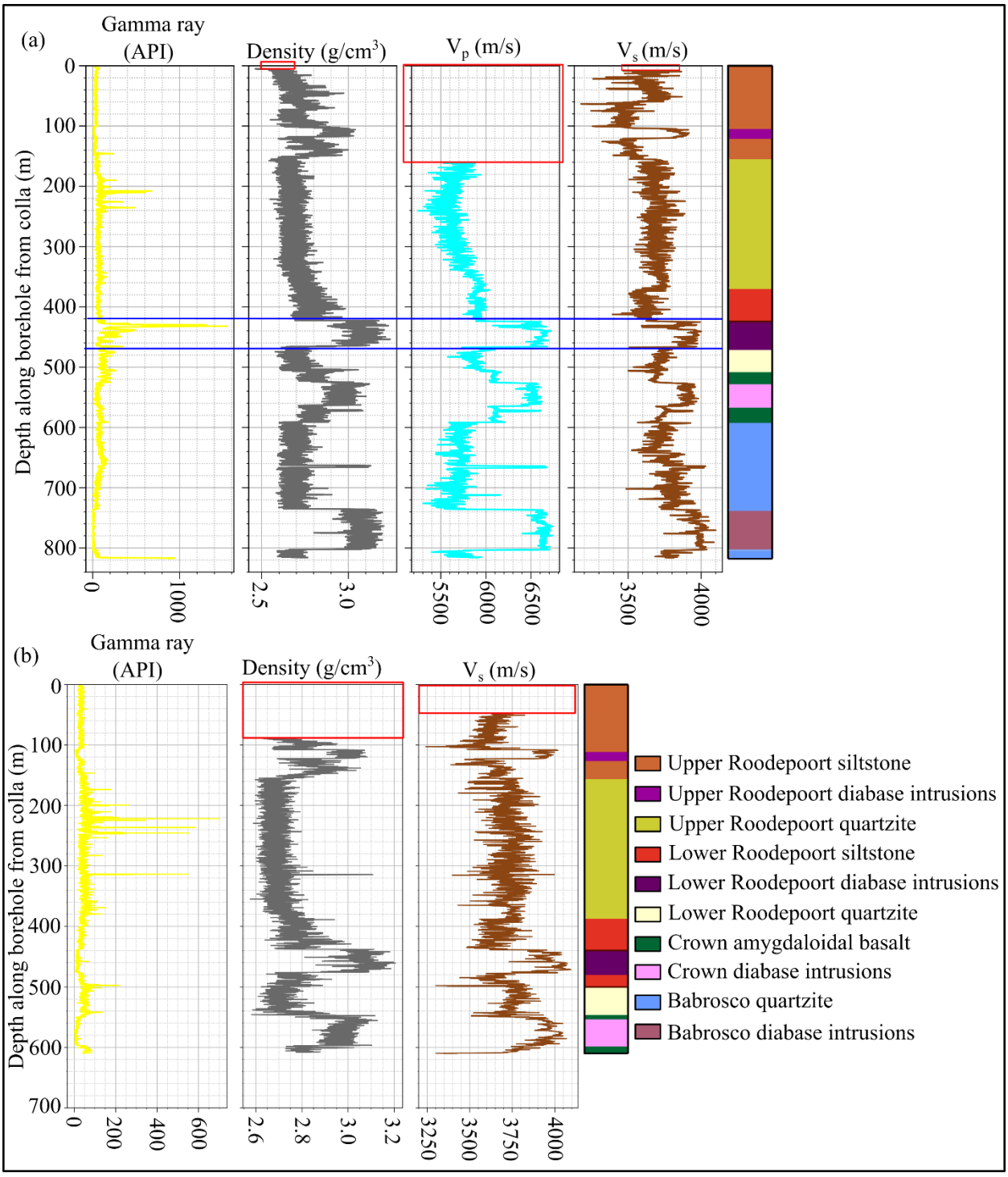

Figure 3.2: Overview of raw geophysical logs with missing data marked in red from ICDP DSeis boreholes at Moab Khotsong gold mine: (a) Borehole A, and (b) Borehole B. Profile sections are colour-coded based on lithology.

3.3.2. Estimation of Borehole B P-wave Velocity Data from Poisson's ratio Estimated from Borehole A Measurements

Poisson's ratio represents the degree of a material's compression in one direction and expansion in the two other directions that are perpendicular to the direction of compression (Kearey and Brookes, 1991). Any material's Poisson's ratio value is a constant solely determined by the material's P- and S-wave velocities. Its mathematical relationship with P- and S-wave velocities is expressed in equation 3.1 (Kearey et al., 2002; Shearer, 1996; Horsfall and Davies; 2019), where σ is Poisson's ratio, V_p is P-wave velocity, and V_S is S-wave velocity.

$$\sigma = \frac{\left(\frac{V_p}{V_s}\right)^2 - 2}{2 * \left(\frac{V_p}{V_s}\right)^2 - 1} \tag{3.1}$$

P-wave velocity data for borehole B from 40- to 610 m was estimated by computing the constant Poisson's ratio for each lithology type in the study area using the V_p and V_s measurements from borehole A using equation 3.1. Subsequently, the computed value of the Poisson's ratio for each lithology class is substituted into equation 3.2 to establish the relationship between P- and S- waves velocities for each lithological class in borehole B. The resulting Poisson's ratio values, V_p/V_S ratio, and the relationship between V_p and V_s for each lithological type are listed **in Table 3.5.**

$$\frac{V_p}{V_S} = \sqrt{\frac{2(1-\sigma)}{(1-2\sigma)}} \tag{3.2}$$

Table 3.5: Poisson's ratio constant, Vp and Vs relationship for each lithology type at Moab Khotsong gold mine DSeis boreholes. UR and LR denote Upper Roodepoort and Lower Roodepoort formations, respectively. N denotes number of data samples.

Lithology type	Poisson's ratio	V_p and V_s relationship (m/s)	V_p and V_s ratio
UR quartzite (N = 4147)	0.13 ± 0.03	$V_p = 1.53V_s$	1.53 ± 0.04
LR metasiltstone (N = 1107)	0.20 ± 0.01	$V_p = 1.63V_s$	1.63 ± 0.02
LR diabase intrusion (N = 942)	0.22 ± 0.02	$V_p = 1.67V_s$	1.67 ± 0.03
LR quartzite (N = 711)	0.17 ± 0.01	$V_p = 1.58V_s$	1.58 ± 0.04
Crown basalt 1 (N = 398)	0.21 ± 0.01	$V_p = 1.65V_s$	1.65 ± 0.01
Crown diabase intrusion (N = 783)	0.22 ± 0.01	$V_p = 1.67V_s$	1.67 ± 0.02
Crown basalt 2 (N = 501)	0.20 ± 0.01	$V_p = 1.64V_s$	1.64 ± 0.02
Babrosco quartzite 1 (N = 2923)	0.11 ± 0.04	$V_p = 1.51V_s$	1.51 ± 0.05
Babrosco diabase intrusion (N = 1323)	0.22 ± 0.01	$V_p = 1.67V_s$	1.67 ± 0.01
Babrosco quartzite 2 (N = 250)	0.10 ± 0.04	$V_p = 1.50V_s$	1.50 ± 0.04

In summary, the Poisson's ratio for the study area ranges from 0.100 ± 0.04 to 0.20 ± 0.01 for the metasediments (quartzites and siltstone), 0.20 ± 0.01 to 0.21 ± 0.01 for Crown basalt, and 0.22 ± 0.01 to 0.22 ± 0.02 for diabase intrusions. The conducted laboratory Poisson's ratio by Nkosi et al. (2022) shows a mean Poisson's ratio of 0.20 for the metasediments, 0.24 for the Crown basalt, and 0.23 for the diabase intrusions. The quartzite layers exhibit the lowest value of Poisson's ratio relative to the intrusive rocks, siltstone, and Crown basalt. The lower Poisson's ratio of the quartzite labels may be due to the higher content of quartz, as Poisson's ratio decreases with increasing quartz content. These results are consistent with the results made by Nkosi et al. (2022). The increased content of plagioclase in the intrusions may be responsible for the highest Poisson's ratio in the diabase intrusions, as it is known that Poisson's ratio increases with increasing plagioclase content. These results also agree with the results made by Nkosi et al. (2022).

V_p and V_S ratio ranges from 1.53 ± 0.04 to 1.58 ± 0.04 for Roodepoort quartzite, 1.63 ± 0.02 for metasiltstone, 1.64 ± 0.04 to 1.65 ± 0.01 for Crown basalts, 1.51 ± 0.05 to 1.50 ± 0.04 for Babrosco quartzite, and 1.67 ± 0.01 for diabase intrusions (**Table 3.5**). The results of V_p and V_s ratio

laboratory measurements conducted by Ishida et al. (2018); and Rickenbacher (2018) on the retrieved DSeis core samples revealed that V_p and V_s ratio is 1.60 ± for Roodepoort quartzite, 1.59 ± 0.02 for metasiltstone, 1.58 ± 0.03 for Crown basalt, 1.57 ± 0.03 for Babrosco quartzite, and 1.58 ± 0.04 to 1.60 ± 0.02 for diabase intrusions.

When our estimate V_p log data plot is compared with V_p log data plot estimated by Nkosi et al. (2022), (**Figure 3.3**), it shows that the observation made by Nkosi et al. (2022) is higher. This may be because Nkosi et al. (2022) used a consistent value of 1.60 as the V_p and V_s ratio for all the lithology types listened in **Table 3.5** (V_p and V_S relationship of $V_p = 1.60V_s$ for all lithologies), while we estimated the V_p and V_s ratio for each lithology type using equation 3.2.

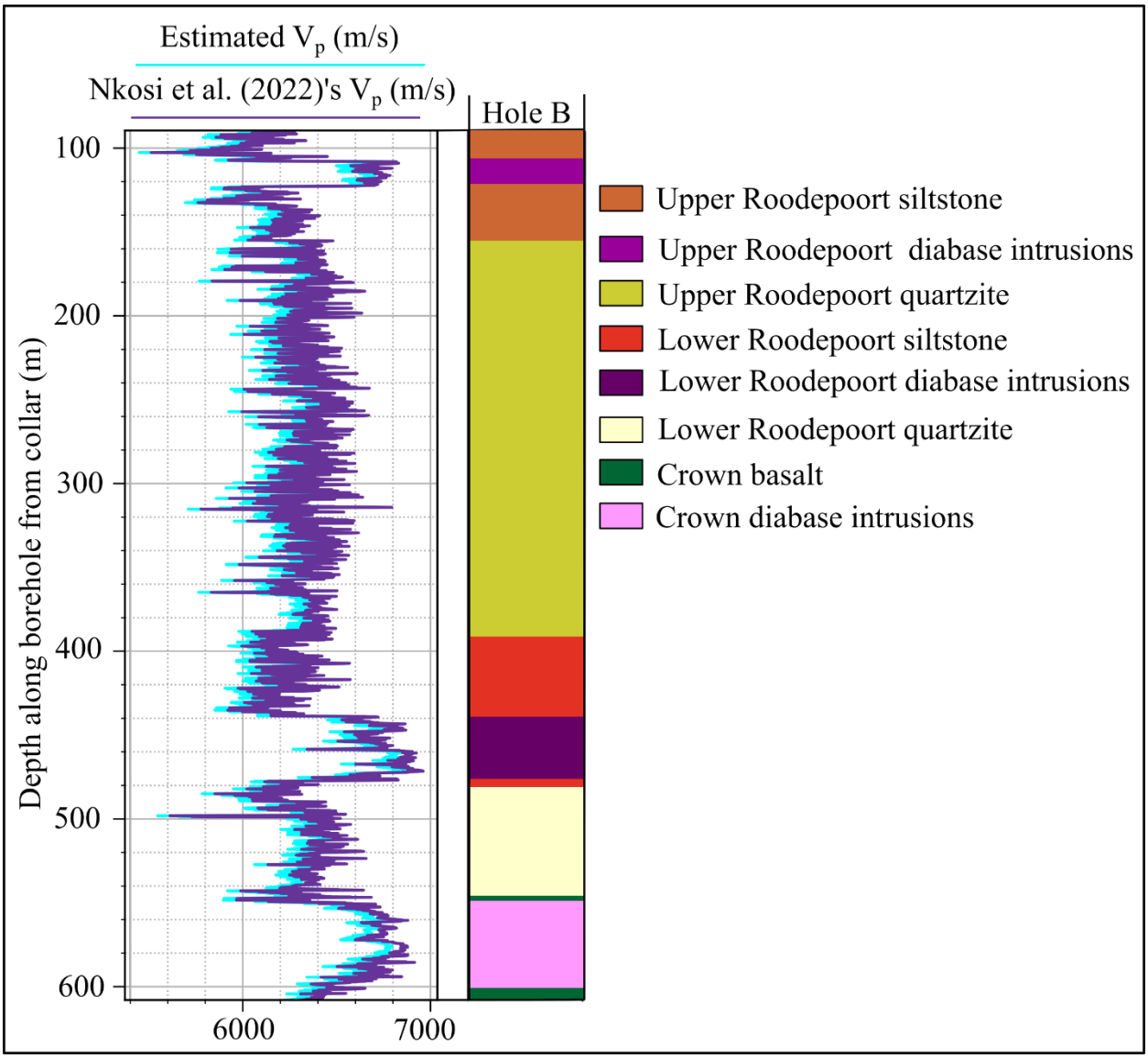

Figure 3.3: Comparison of our estimate V_p log plot (marked in cyan colour) with Nkosi et al. (2022)'s observed V_p log plot (marked in purple colour)

4. LITHOLOGY CLASSIFICATION METHODOLOGY

4.1. Overview of Machine Learning-based Algorithms

Machine learning is a component of Artificial Intelligence (AI), where a computer is trained and made to perform task(s) from its prior knowledge using statistical models and algorithms. With the availability of large quality datasets and user-friendly software packages (e.g., Python), the demand for machine learning is high and many machine learning algorithms have been widely adopted.

Supervised and unsupervised algorithms are two main categories of machine learning-based algorithms. Supervised machine learning algorithms are applied to infer based on labelled dataset. The input dataset is split into training and testing sets. Supervised algorithms make prediction from the testing dataset applying patterns learnt from the training dataset (Mahesh, 2020). Examples of common supervised machine learning algorithm include gradient boosting decision trees, neural network, random forest, and support vector machine.

Unsupervised machine learning algorithms are based on an unlabelled dataset. The algorithms create subsets or clusters within the dataset based on the similarity or dissimilarity of characteristics represented by each data feature to form a classification scheme (Lofts, 1993; Rabaute et al., 1997; Mahmoodi and Smith, 2015; Mahesh, 2020). Then the classification scheme is used to develop clusters from fresh dataset. Examples include hierarchical and non-hierarchical cluster analysis.

Factors to consider when choosing machine learning algorithms include computational complexity, model complexity, model performance characteristics, robustness to prediction bias and outliers, data density, input variance, input and prediction noise, and features interactions (Nwaila et al., 2022).

4.1.1. Classification Models

The below discussed machine learning classification algorithms were chosen because they are well-known techniques widely applied in data analysis procedures to identify formation lithology in the geosciences due to their high performance characteristics in lithology classification and robustness to prediction bias, outliers, imbalance data, and small datasets (e.g., Xie et al., 2018; Sun et al., 2019; Zhang et al., 2021; Zou et al., 2021; Nwaila et al., 2022). To the best of our knowledge, no publication on the application of these algorithms to lithology classification in the study area has been done.

4.1.1.1. Random Forest (RF) Algorithm

Breiman first proposed the Random Forest (RF) algorithm, an ensemble supervised machine learning algorithm based on the random construction of several uncorrelated decision trees (Breiman et al., 1984). A decision tree is a binary tree structure. This technique significantly reduces computational time by continually dividing the dataset into two categories for either regression or classification.

According to Breiman (2001), for classification problems, a decision tree is constructed by splitting randomly on a selected feature from the training dataset that has the maximum information gain and minimum information entropy to be the root node (information gain is a measure of how much information a given feature in a dataset contains, and information entropy is a measure of uncertainty in a dataset). This process is repeated for the remaining features by consistently splitting on the feature that has the maximum information gain and minimum information entropy randomly. Then, it is added as a child node of the root node. The same process is repeated several times until the tree reaches the desired great size to build the forest (Aha and Breslow, 1998; Xie et al., 2018). The final prediction result is the majority voted prediction given by the individual decision trees (**Figure 4.1**).

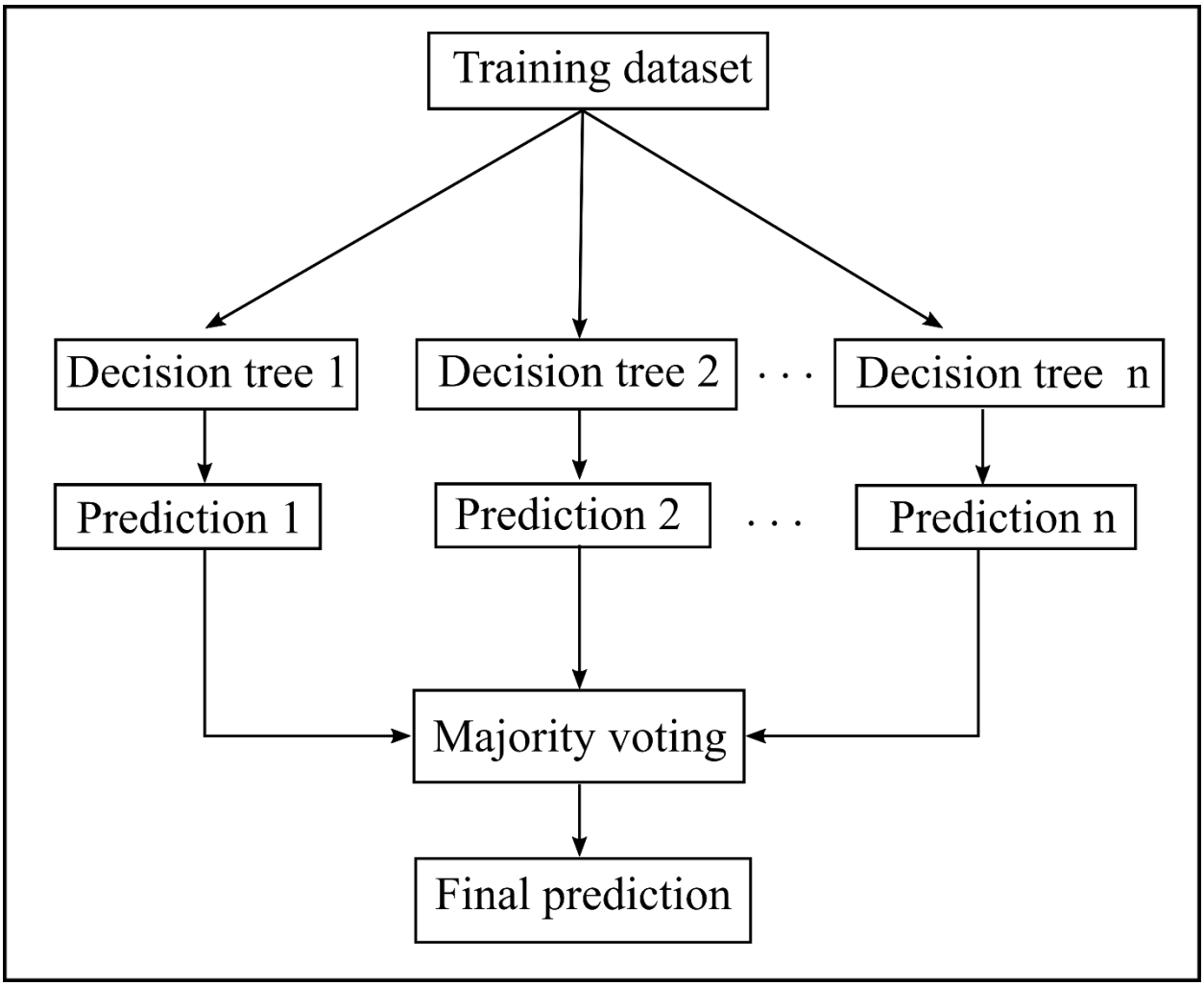

Figure 4.1: Graphical representation of Random Forest algorithm (from Breiman, 2001).

Given a training dataset $x = x_1 \ldots x_n$ with responses $y = y_1 \ldots y_n$

$$for\ i = 1 \ldots N$$

Where n is dimension of training dataset and i is number of trees. Sample randomly with replacement from x and y, call these x_i, y_i. Train a classification model on x_i, y_i and the final prediction is the majority vote over the forest (Xie et al., 2018).

The RF algorithm has been applied frequently in engineering studies (Zhao et al., 2014; Grinand et al., 2013; Li et al., 2014; Attarchi and Gloaguen, 2014), but has rarely been applied in the

classification of lithology from mineral deposits. Of recent, Cracknell and Reading (2014); Kuhn et al. (2018); Xie et al. (2018); Sun et al. (2019); Zou et al. (2021) have accurately applied the RF technique to improve the classification of lithology based on geophysical logs dataset from mineral deposits.

The specific procedures in Random Forest algorithm are summarized as follows (Sun et al., 2019):

1) Random selection of new sample set and construction of k decision trees from the training data by applying the Bootstrap method (i.e., random selection of sample from training data with replacement),
2) Determination of the amount of information each feature contained, and selection of the feature with significant classification potential to split a node,
3) Allow maximum growth of each decision tree, and
4) Composition of the Random Forest classifier by the several generated decision trees, and the application of the classifier to evaluate and classify new data (test datasets). The final classification outcome is the most voted prediction result given by the decision trees.

The RF classification algorithm has several advantages over other known machine learning-based classification algorithms, for example, single classifiers. Advantages include better performance accuracy by applying the concept of 'Bagging', where results of several decision tree are combined to produce consistent result with reduced variance, less susceptibility to overfitting (situation where the training and testing accuracy scores are too far apart e.g., a difference > 20%, causing the generalization ability of the classifier model to be low, and can be avoided by optimizing the hyperparameters) due to random sampling and significant improvement in the speed in training samples of large dataset (Breiman,1996; Parmar et al., 2014; Sun et al., 2019). Also, the RF algorithm is made more robust by applying the parameter Bootstrap. The concept of Bootstrapping is to ensure subset of the total training dataset is worked on by each tree (Breiman, 1996; Parmar et al., 2014) and can be applied to high dimensional data. However, the RF model has the disadvantage of making poor predictions with small or low-dimensional data (Sun et al., 2019). The hyperparameters of RF model is described in **Table 4.1**.

Table 4.1: Hyperparameters for each supervised classifier model

Supervised model	Hyperparameters
Random Forest (RF)	number of estimators (the number of trees required to build the forest). maximum depth (the maximum growth depth of each tree). minimum samples split (the minimum number of samples an internal node required to split).
Support Vector Machine (SVM)	C (Penalty parameter that controls misclassification of data points). Gamma (the spread of the kernel function).
Gradient Boosting Decision Trees (GBDT)	learning rate (the rate at which the contribution of each tree shrinks). number of estimators (the number of boosting stages). maximum depth (the maximum growth depth of each tree). minimum samples split (the minimum number of samples an internal node required to split). Minimum samples leaf (the minimum number of samples required at a leaf node).

4.1.1.2. Gradient Boosting Decision Tree (GBDT)

GBDT is a boosting ensemble supervised learning algorithms proposed by Friedman (2002). GBDT builds a decision tree based on the combination of several weak decision trees, which is generated in a sequence by reducing the combined decision tree bias at each step (Friedman, 2002; Xie et al., 2018). Examples of GBDT applications include classification and identification of natural minerals (Akkaş et al., 2015), and lithology identification using geophysical data (Han et al., 2018; Bressan et al., 2020). The GBDT model tends to establish a weak decision tree using the training dataset and the fitting residual error is then calculated. A new weak decision tree is established by learning the loss function's negative gradient of the former weak decision tree. This eventually results in the establishment of a strong decision tree by combing a large number of weak decision trees as seen in **Figure 4.2**, thus minimizing the loss function by reducing the residual error value gradually (Jia and Xue, 2022). The applied hyperparameters of GBDT are listed in **Table 4.1**.

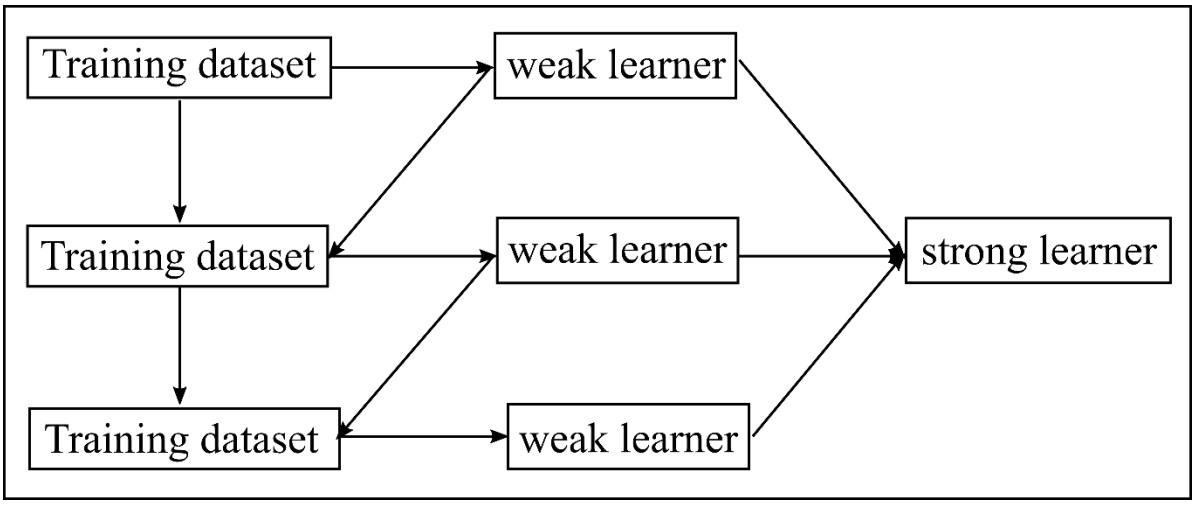

Figure 4.2: Graphical representation of gradient boost decision tree algorithm (from Friedman, 2002).

Given a training dataset N = $\{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$, setting the loss function $L(y, f(x)) = -\sum_{k=1}^{K} y_k log p_k(x)$, where x is the feature vector of dimension n, y is the target and $p_k(x) = \exp(f_k(x)) / \sum_{l=1}^{k} \exp(f_l(x))$ is the probability of class k, a weak classifier $f_0(x)$ implemented by a regression tree is initialized to minimize the loss function $L(y, f(x))$ as seen in equation 4.1 (Friedman 2002; Jia and Xue, 2022; Zou et al., 2021), where r is the negative gradient value of the loss function.

$$f_0(x) = argmin \sum_{i=1}^{N} L(y_{i,} r) \tag{4.1}$$

For each iteration of the trained regression tree, i.e., the m-th iteration, the negative gradient value of the loss function (r) of the current classifier for sample i is calculated, as shown in equation 4.2 (Jia and Xue, 2022; Zou et al., 2021):

$$r_{mi} = -\left[\frac{\partial L\left(y_{i,} f(x_i)\right)}{\partial f(x_i)}\right]_{f(x) = f_{m-1}(x)}, i = 1, \dots, n \tag{4.2}$$

For the m-th regression tree, the j-th parameter leaf node is calculated as equation 4.3:

$$r_{mj} = argmin \sum_{x_i \in R_{mj}} L\left(y_i, f_{m-1}(x_i) + rI\left(x \in R_{mj}\right)\right), j = 1, \ldots, J \quad 4.3$$

where R_{mj} represents the region associated with the j-th leaf node in the m-th regression tree. Update the classifier according to different parameter learning rate η as shown in equation 4.4:

$$f_m(x) = f_{m-1}(x) + \eta \sum_{j=1}^{J} r_{mj} I\left(x \in R_{mj}\right) \quad 4.4$$

A final strong classifier is obtained by combining M updates of weak classifiers after M iterations according to different learning rates η as follows (equation 4.5):

$$f_M(x) = f_0(x) + \sum_{m=1}^{M} \eta \sum_{j=1}^{J} r_{mj} I\left(x \in R_{mj}\right) \quad 4.5$$

The GBDT algorithm has the advantage of being robust to outliers and missing values like every other ensemble method and can outperform random forest model if properly tuned (Maxwell et al., 2019). The key disadvantage of GBDT is that it can be prone to overfitting in a case of high learning rate and high number of trees (Maxwell et al., 2019; Zou et al., 2021) and computationally expensive because of many hyperparameters (Maxwell et al., 2019).

4.1.1.3. Support Vector Machine (SVM)

The SVM is a supervised machine learning algorithm based on statistical learning theory proposed by Vapnik (1995) applicable for regression and classification problems. Support Vector Machine as classifier are maximum dual margin models applicable on linearly and nonlinearly separable datasets (**Table 4.1**). The SVM classifier differentiates classes by building hyperplane(s), which depend on the patterns and relationship between the datapoints. Support vectors are data points nearest to the hyperplane line, and margin is the distance between these support vectors (**Figure 4.3**). In general, the Support Vector Machine (SVM) model aims to find the support vector that best characterizes the hyperplane with the most maximized margin, which is then believed to be the most robust model with minimum generalization error rate. SVM has been proposed to show superior performance in nonlinear classification and pattern recognition problems (Hsu and Lin,

2002; Bressan et al., 2020; Al-Anazi and Gated, 2010). Al-Anazi and Gated (2010) and Nwaila et al.(2022) both successfully applied Support Vector Machine (SVM) to classify lithofacies.

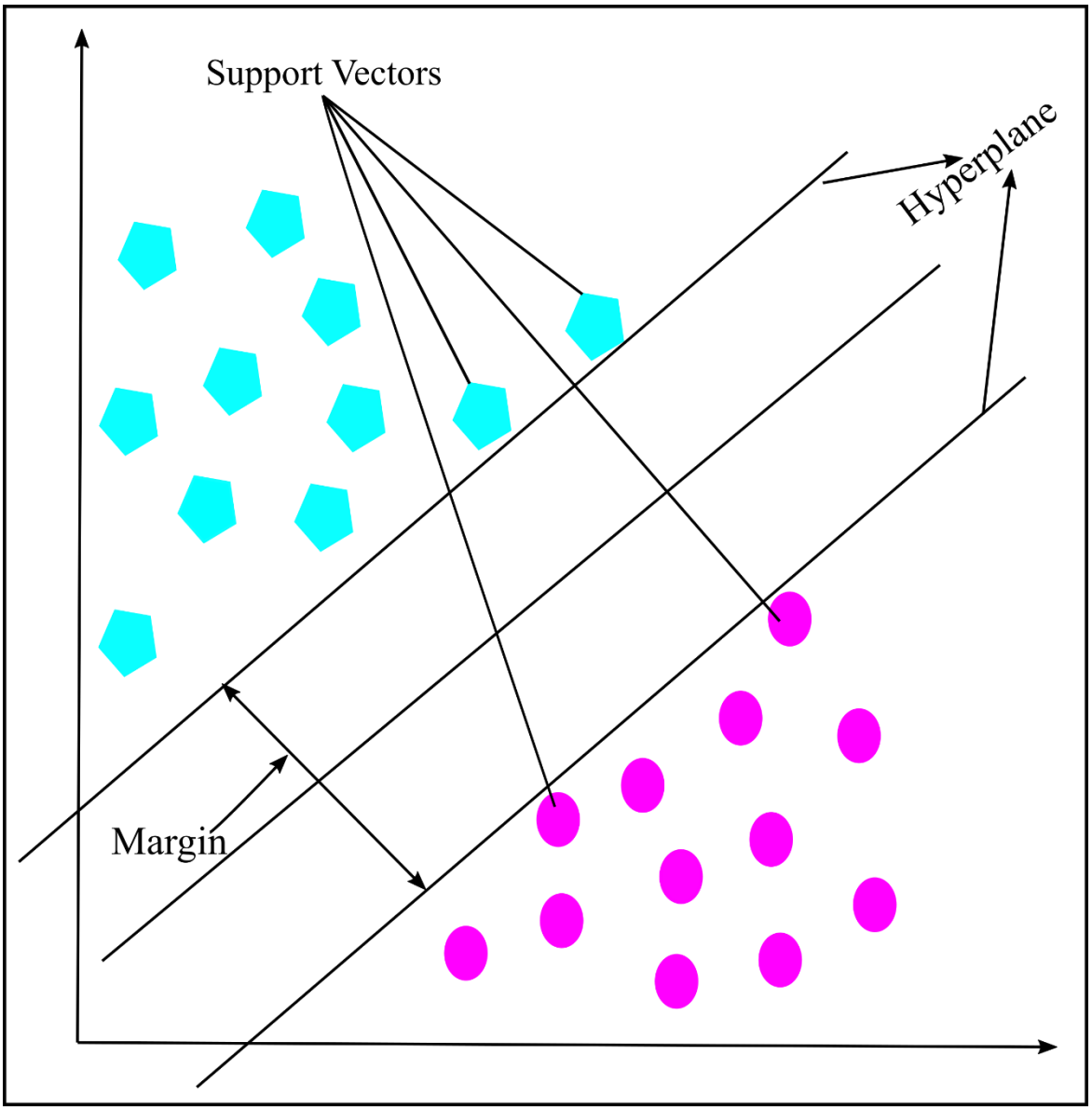

Figure 4.3: Graphical representation of support vector machine algorithm (from Vapnik, 1995).

For linearly separable support vector classifier model, given a training dataset $x_i \in R^n$, where i = (1, …, N), n is the feature vector dimension of the input measurement space, N is the total number of training datapoints, and the binary classification is given as $y_i \in \{-1, 1\}$, in the high

dimensional space, during training, the indicator function is stated in equation **4.6**, where w is the hyperplane normal weight vector that regularizes the classification and b is a scalar bias.

$$f(x) = sign(w^T x_i + b) \tag{4.6}$$

The above linearly separable classifier (equation **4.6**) can be extended to accommodate dataset overlapping by adding non-negative slack variable $\xi_i, i = 1, \dots N,$ as follows (Kecman, 2005):

$$\text{Maximize} \quad \frac{1}{2} w^T w + C \sum_{i=1}^{N} \xi_i$$

$$\text{Subject to} \begin{cases} w^T x_i + b \geq 1 + \xi_i & \text{for } y_i = 1, \xi_i \geq 0, i = 1, \dots, N \\ w^T x_i + b \leq -1 + \xi_i & \text{for } y_i = -1, \xi_i \geq 0, i = 1, \dots, N \end{cases} \tag{4.7}$$

Minimizing $\frac{1}{2} w^T w$ (regularization term) means maximizing the margin between two groups of data (Al-Anazi and Gated, 2010), where C is a basic hyperparameter of SVM called the penalty parameter (see **Table 4.1**). C controls the degree of misclassification errors in training data points. In practice, the optimal value of parameter C is selected from a large array of values through gridsearch and K-fold cross-validation methods using only the training dataset. When implementing the K-fold cross-validation method, the training dataset is divided into K subsets, where one subset is used for validation and the rest for training. This process is repeated K times to ensure every subset is used as the validation subset, and the value of C that has the minimum error averages over the validating subsets is the optimal value (Al-Anazi and Gated, 2010; Xie et al., 2018).

The dual problem of equation **4.7** with inequality constraints can be solved using the Lagrange optimization function below (Kecman, 2005):

$$L(w, b, \xi, \alpha_i \beta_i) = \frac{1}{2} w^T w + C \sum_{i=1}^{N} \xi_i - \sum_{i=1}^{N} \alpha_i \left(y_i (w^T x_i + b) - 1 + \xi_i \right) - \sum_{i=1}^{N} \beta_i \xi_i \tag{4.8}$$

where α_i and β_i are Lagrange multipliers. The Lagrangian function L must be minimized and maximized with respect to w, b, ξ and the non-negative Lagrange multipliers, respectively, to

search for the optimum point. The solution of the problem in equation **4.8** can be achieved either in the primary space of w, b and ξ or in the dual space of the Langrange multipliers (Al-Anazi and Gated, 2010). Kecman (2005) proposed the Karush-Kuhn-Tucker (KKT) conditions as the appropriate conditions to maximize equation **4.8** with respect to α_i. At the optimum point, these conditions represented by the subscript 'o' are as follows:

$$\frac{\partial L}{\partial w_o} = 0 \Rightarrow w_o = \sum_{i=1}^{N} \alpha_i y_i x_i \tag{4.9}$$

$$\frac{\partial L}{\partial b_o} = 0 \Rightarrow \sum_{i=1}^{N} \alpha_i y_i = 0 \tag{4.10}$$

$$\frac{\partial L}{\partial \xi_{io}} = 0 \Rightarrow \alpha_i + \beta_i = C \tag{4.11}$$

The following complementary KKT conditions must be met also:

$$\alpha_I(y_i(w^T x_i + b) - 1 + \xi_i) = 0, i = 1, \dots, N \tag{4.12}$$

$$\beta_i \xi_i = (C - \alpha_i) = 0, i = 1, \dots N \tag{4.13}$$

Substituting equations: **4.9**, **4.10** and **4.11** into equation **4.8** yields the following dual Lagrangian variables:

$$L_d(\alpha) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{N} y_i y_j \alpha_i \alpha_j x^T{}_i x_j \tag{4.14}$$

where $\alpha = [\alpha_1, \dots \alpha_N]^T$. Maximizing (**4.14**) with respect to α_i, subject to the following below constraints yield the optimal hyperplane:

$$\begin{cases} \alpha_i \geq 0, i = 1, \dots, N \\ \sum_{i=1}^{N} \alpha_i y_i = 0 \end{cases} \quad \text{for a separable linear classifier} \tag{4.15}$$

$$\begin{cases} C \geq \alpha_I \geq 0, i = 1, \dots N \\ \sum_{I=1}^{N} \alpha_i \, y_i = 0 \end{cases} \quad \text{for an overlapping linear classifier} \qquad 4.16$$

For a model with nonlinearly separable support vectors, kernel functions are required to transform the input data $x \in R^n$ dimensional space into data $\Phi(x)$ of a higher dimensional space known as feature space, where the support vectors can be separated using the linear classifier formulation demonstrated above in equation **4.14**. The transformation to feature space is the major idea that controls the separation of nonlinear support vectors (Hamel, 2009; Maji et al., 2013; Gullo, 2015; Sebtosheikh and Salehi, 2015; Ortegon et al., 2018; Zoppis et al., 2019; Al-Mejibli et al., 2020). This transform approach leads to solving a quadratic programming optimization problem like equation **4.14** (Kecman, 2005). The indicator function is reformulated as (Al-Anazi and Gated, 2010):

$$i_F = sign\left(\sum_{i=1}^{N} y_i \alpha_i \, \Phi^T(x_i)\Phi(x_j)\right)$$

$$= sign\left(\sum_{i=1}^{N} v_i K(x_i, x_j) + b\right) \qquad (4.17)$$

where v_i is the weights and $K(x_i, x_j)$ is the kernel function.

For this study, a polynomial kernel (Radial Basis Function (RBF kernel)) was used to build a nonlinear SVM classifier model (equation **4.18**). This is because a polynomial kernel function has been found by several researchers to produce minimum misclassification rate error for nonlinear classification problem (Sebtosheikh and Salehi, 2015; Xie et al., 2018). The RBF kernel or Gaussian kernel can be expressed mathematically as (Al-Mejibli et al., 2020; Sun et al., 2019):

$$K(x_i, x_j) = exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right) \qquad (4.18)$$

where $\frac{1}{\sigma^2}$ represent the hyperparameter Gamma (γ) that controls the spread of the kernel function (see **Table 4.1**). Nonlinear SVM classifier formulation is a generalization on the linear SVM (equation **4.14**), and it is as follows (Al-Anazi and Gated, 2010):

$$\text{Maximize } L_d(\alpha) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{N} \alpha_i \alpha_j y_i y_j K(x_i x_j) \tag{4.19}$$

After the Lagrangian multiplier are computed from optimization of the training data, the indicator function is from the support vectors:

$$i_F = sign\left\{ \sum_{i}^{N} \alpha_i y_i K(x_i x_j) + b \right\} \tag{4.20}$$

Originally, the SVM algorithm was developed to be applied in binary classification. Research on the effective extension of SVM for multiclass classification is still ongoing. To extend the SVM classifier for multi-class classification consisting of l classes, a one-against-one method was proposed by Hsu and Lin (2002), where $\frac{l(l-1)}{2}$ binary SVM model classifiers are constructed, and each classifier is trained on patterns from two classes. Then, the prediction of a class is made based on the highest performance score of the binary classifier on that class relative to all other classes. Hence, with the same data size, it is more computationally costly to solve a multi-class problem than a binary problem. In this study, the multi-lithological classes of the logs were classified using the one-against-one method. The SVM model can be robust to small or biased data if the hyperparameters are well tuned. One of the key limitations of SVM is its sensitivity to outliers.

4.1.1.4. Cluster Analysis

Cluster analysis is a kind of unsupervised multivariate statistical method that employs a series of techniques to generate classes within a dataset according to similarity or dissimilarity features represented by the dataset. Cluster analysis uses the generated class(es) to build a classification scheme that is applied to classify new dataset (Hastie et al., 2009). The possibility of clustering classes of lithologies to map radioactive zone using machine learning was examined by K-means

clustering analysis. The K-means clustering is implemented to further assess the accuracy of machine learning models to classify lithology where the geological knowledge from geologists or geophysicists is poor or absent. In addition, the K-means clustering algorithm may create meaningful similar clusters within available data described by the cluster centroids, aiding detailed predictions and insight by minimizing the within-cluster sum of squared distance (i.e., variance) according to its Euclidean distance (equation **4.21**). In this algorithm, initial cluster centroid with localized minimum within-cluster Euclidean distance is iteratively explored by moving datapoints from one cluster to another.

$$arg\ \min_{S} \sum_{i-1}^{k} \sum_{x \epsilon S_i} \|\chi - \mu_i\|^2 \tag{4.21}$$

where χ is a cluster and μ_i is the cluster mean (also known as cluster centroid). At each iteration, the algorithm performs two steps: firstly, an assignment step where the algorithm groups the datapoints into a random cluster using an initial cluster centroid that minimizes its Euclidean distance using (**equation 4.22**); then, an update step where new cluster centroids are recalculated for the present cluster using (equation **4.23**), (Forgy, 1965). These steps keep iterating until the algorithm converges to global minimum where there are no changes to the centroids and grouping of datasets into clusters is not changing.

$$S_i^{(t)} = \left\{ x_p : \left\| x_p - m_i^{(t)} \right\|^2 \leq \left\| \chi_p - m_j^{(t)} \right\|^2 \forall j, 1 \leq j \leq k \right\}, \tag{4.22}$$

$$m_i^{(t+1)} = \frac{1}{\left| S_i^{(t)} \right|} \sum_{x_{j \epsilon S_i^{(t)}}} x_j \,, \tag{4.23}$$

where x_p is each dataset, $m_i^{(t)}$ are initial cluster centroids at each iteration "t", $S_i^{(t)}$is the new cluster generated by assigning each dataset with the minimum Euclidean distance, and $m_i^{(t+1)}$ is the recalculated centroids for the new generated clusters.

4.1.2. Code Availability

The supervised and unsupervised machine learning-based models were implemented from the SciKit-Learn module in Python programming language using the Jupyter Notebook in Anaconda software. Code available at https://github.com/obehi-atita/Lithology-Classification-using-ML Hardware required is windows10, core i7 with 32 GB memory RAM. This study utilizes "SuperVectorClassifier", "RandomForestClassifier" and "GradientBoostingClassifier" from SciKit-learning module to generate super vector machine model, Random Forest model, and gradient boosting decision tree model, respectively (Pedregosa et al., 2011; Zou et al., 2021). The Strater 5 logging software was used to plot geophysical borehole data and lithological logs https://d.docs.live.net/99bc8b947bc00b38/hole%20raw_plots_strater.sdg

For supervised classification methods, we employed six major stages of machine learning workflow modified after Zhang et al. (2021), which includes (**Figure 4.4**):

1) Data preprocess,
2) Data splitting into training and testing datasets,
3) Optimization and hyperparameter tuning,
4) Model building for classification of lithologies using the training dataset,
5) Prediction performance evaluation, and
6) Analysis and verification of predicted lithologies.

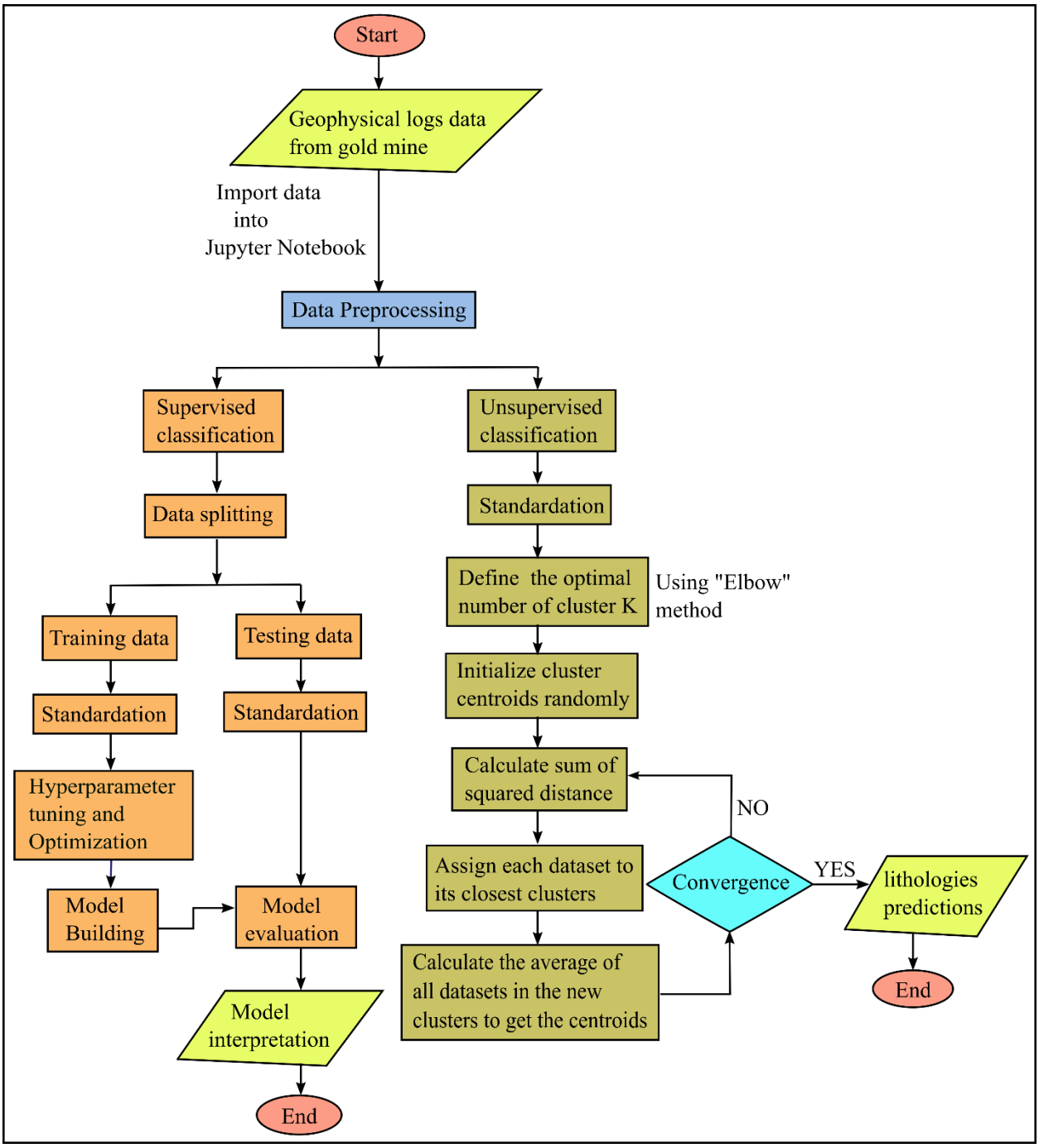

Figure 4.4: Summarized workflow of proposed applied machine learning-based methods.

For unsupervised classification methods, we employed two major stages of machine learning workflow modified after Zhang et al. (2021) which includes (**Figure 4.4**):

1) Data preprocessing, and
2) Selection of optimal number of clusters.

4.1.3. Data Preprocessing

The data preprocessing step is essential to make the log dataset suitable for the implementation of machine learning algorithms for the prediction of lithologies. The four logging variables (short spaced bulk density, natural total gamma-ray, V_p and V_s) made available by the DSeis-ICDP project team were combined for the implementation of our machine learning models (**Figure 3.2**). A total of 20153 log data points from borehole A and a total of 13741 log data points from borehole B were provided. The tools used in the geophysical logging are discussed in section **3.3**, and the processing and spacing interval of the log data is discussed in chapter 3.3.1, based on the information provided by members of the ICDP DSeis team in Japan.

For supervised classification, combination of ten classes of lithologies (Upper Roodepoort quartzite, Upper Roodepoort siltstone, Upper Roodepoort diabase intrusion, Lower Roodepoort siltstone, lower Roodepoort quartzite, lower Roodepoort diabase intrusion, Crown basalt, Crown diabase intrusion, Babrosco quartzite and Babrosco diabase intrusion), (see **It** is important to accurately classify these lithologies between layers at IDCP DSeis boreholes situated at the Jeppestown subgroup of the West Rand Group of Witwatersrand Supergroup for safety mapping and objective decision making.

Table 2.1), and four log features (short-spaced bulk density, natural gamma ray, V_p, and V_s) were used (see **Figure 3.2**). The lithostratigraphic classes are encoded as shown in **Table 4.2** for easier data analysis using machine learning algorithms. These collected datasets were transformed into a matrix of rows and columns comprising the above mentioned logging variables and labelled

classes of lithostratigraphies from core logging report (see a sample in **Table 4.3**). For unsupervised classification, only the four logging features were used, as labelled data are not required

Table 4.2. Encoded classes of lithostratigraphy from DSeis boreholes A and B.

Hole A			Hole B		
Lithostratigraphy Code	Description	Data size	Lithostratigraphy Code	Description	Data size
0	Upper Roodepoort quartzite	4147	0	Upper Roodepoort quartzite	4721
1	Lower Roodepoort siltstone	1107	1	Upper Roodepoort siltstone	1012
2	Lower Roodepoort diabase intrusion	942	2	Upper Roodepoort diabase intrusion	300
3	Lower Roodepoort quartzite	711	3	Lower Roodepoort siltstone	1059
4	Crown basalt	898	4	Lower Roodepoort diabase intrusion	741
5	Crown diabase intrusion	783	5	Lower Roodepoort quartzite	1300
6	Babrosco quartzite	3183	6	Crown basalt	200
7	Babrosco diabase intrusion	1323	7	Crown diabase intrusion	1040

Table 4.3: A sample of log dataset structured into a matrix.

Gamma ray (API)	Density (g/cm^3)	V_p (m/s)	V_s (m/s)	Lithology
45.07	2.63	5795.98	3695.04	0
47.98	2.65	5783.32	3701.42	0
49.20	2.68	5798.97	3680.98	0
52.29	2.63	5769.23	3671.97	0
59.32	2.67	5742.73	3671.52	0
…	…	…	…	…
59.74	2.63	5804.06	3811.94	6
58.44	2.64	5810.35	3792.67	6
61.13	2.69	5832.94	3769.32	6
68.62	2.70	5886.97	3768.84	6
77.94	2.73	5915.76	3754.69	6

To make the log dataset suitable for the implementation of machine learning-based algorithms to achieve our aim of lithology classification, the following basic machine learning preprocessing steps were carried out on the log datasets:

4.1.3.1. Missing Log Data Values

Missing data values can be described as those values that are not recorded for a feature in a dataset. Missing data is commonly observed basically in all surveys, regardless of how well planned and controlled the surveys were, and how carefully the data were acquired. Missing data can cause the reliability of statistical analysis of a dataset to be reduced, unreliable parameter estimates, and the analysis of the whole dataset to be complex. Consequently, these significant impacts on a dataset can lead to invalid conclusions from the dataset (Kang, 2013). So, it is important to deal with any missing data values in a dataset before conducting a full analysis of the complete dataset.

There are numerous techniques for handling missing data value in a dataset. Techniques include:

1) Listwise or case deletion, where the whole case sections with missing data are completely removed (Kim and Curry, 1977; Graham, 2009; Kang, 2013). For example in **Figure 3.2a**, all columns of borehole A data section where the V_p feature has missing values from 0-180 m will be removed. This may result in the loss of significant information in the dataset and reduction in the data size, which can negatively impact the accuracy of the models,
2) Mean/Median substitution, where the missing data value in a feature is substituted with the estimated mean/median value for that same feature (Kang, 2013). This minimizes loss of information. However, it can increase outliers' impact in the case of mean imputation. Also, it can distort the original distribution and pattern of the dataset and reduces the data variance. Consequently, this can introduce prediction bias into the models,
3) Last observation carried forward, where a missing data value is replaced with the last observed value (Hamer and Simpson, 2009; Kang, 2013). This can introduce outliers in the dataset and change the original pattern of the dataset. This also can cause the models' prediction to be biased,
4) Regression imputation, where the missing data value is replaced with an estimated value from linear or logistic regression model fit using available data from dependent variable (Kang, 2013). This method takes into consideration the relationship between variables containing the missing values and other variables. However, it can be affected by outliers

and does capture non-linear relationships in the dataset, which can affect the performance of the models negatively,

5) Linear interpolation, where existing recorded values after and before the missing value are used to estimate the missing data value etc. This method was applied to estimate the missing data value in this study. This method of handling missing values preserves variance and pattern in the dataset. The disadvantages of this method include incorrect results if missing values are consecutive, which may result in poor prediction performance of the models.

Figure 3.2 shows missing data values, highlighted in colour red, in borehole A from collar distance of 0-160 m and 815.77-817.24 m in the V_p dataset; from 1-8 m and from 815.67-817.24 m in the V_s dataset; from 1-8 m and from 815.25-817.24 m in the short-spaced bulk density dataset, while the gamma ray dataset is complete. Similarly, data values are missing in borehole B from collar distance of about 1-89.39 m and 607-610.04 m in the short-spaced bulk density dataset; 1-49.02 m in the V_s dataset; and the whole of V_p dataset. To successfully analyze the downhole geophysical logged dataset combined, the features must have the same data size. Hence, the core section of gamma ray, bulk density and V_s log data values that corresponds with the large segments of missing data values far outside recorded known values (e.g., from 0 – 160 m core section of V_p data in borehole A) were removed. All the logging features in borehole A were sorted to sampled collar depth of V_p (161.01-815.77m) using a spacing interval of 0.05m. Consequently, this reduced the dataset size to 13091 and the amount of possible available information that might be present in the log datasets. The logging features in Hole B were sorted from collar depth of 89.39-607.98 m using a spacing interval of 0.05 m to make a total data size of 10373. However, individual missing data value within recorded known data values for the logs (**Table 4.4**) were estimated and imputed using the linear interpolation technique formula given below (Chapra and Canale, 1998):

$$y = y1 + \frac{(x - x1)(y2 - y1)}{(x2 - x1)} \qquad (4.24)$$

where $x1$ and $y1$ are the first data values above missing value, $x2$ and $y2$ are the first data values below missing value, x is the independent data value at the point to perform the interpolation, and

y is the dependent interpolated data value. Due to the constant variation of logging features in the dataset depending on the lithology type, using recorded known data values above and below the missing data values to estimate the missing value has advantage of retaining the original distribution of the feature being interpolated because it preserves the variability and patterns in the dataset. In this regard, all interpolated data point values (2200 data points in total in borehole A and 1300 in borehole B) had maximum deviations of about $\pm$ 0.01.

Table 4.4: Missing data values within recorded known values in boreholes A and B.

Hole A		Hole B	
Logging features	Missing data points	Logging features	Missing data points
Gamma ray	0	Gamma ray	0
Density	400	Density	500
V_p	800	V_p	-
V_s	1000	V_s	800

4.1.3.2. Detection and Removal of Outliers

An outlier is a data value that differs significantly from the range of other data values present in a dataset to such a degree that there is a concern that it might have been generated by a different process, for example, cycle-skipping in sonic logs, which means that the first break is not picked because of poor signal to noise (Hawkins, 1980; Dan and Ijeoma, 2013). In a dataset, outliers may occur due to several causes. These causes can be grouped into two main classes (Anscombe, 1960): (1) those that occur from errors, for example errors in data acquisition and recording, and (2) those that occur from the natural variability within the dataset.

The presence of outliers in a dataset can cause statistical analysis based on means, standard deviation and variance to be unreliable. Thus, can lead to an invalid conclusion from the dataset. This is because mean and variance are highly sensitive to outliers (Zimmerman, 1994; Dan and Ijeoma, 2013). In addition, the presence of outliers can significantly affect the performance of most machine learning algorithms. However, not all outliers are bad. Some could provide insight for

further studies. Therefore, there is a need for a proper investigation before an outlier is being removed.

In order to enhance the classification performance of our proposed implemented machine learning-based methods, we had to identify and impute some of the outliers in our applied dataset. In this study, outliers may be described as those data points that are noticeably distinct from the overall pattern of other data points within a lithological layer of a log dataset. Possible outliers in the geophysical log dataset were detected by first visualizing each logging feature for every lithological type by plotting each log against borehole depth to understand data pattern (**Figure 4.5Figure 4.6**). Outliers were further identified by plotting a boxplot for each log feature for every lithology (**Figure 4.6**). The center line in the coloured box demonstrates the median for the data. If the data are symmetrical, the median will be in the center of the box (for example, **Figure 4.6b**). If the data are skewed, the median will be closer to the upper or lower of the coloured box (for example, **Figure 4.6a**). The upper and lower boundaries of the coloured box show the first and third quartiles, respectively. Lines that extend from the box are whiskers, which show the expected variation of each log data. The top and bottom end of the whiskers represent the maximum (Max) and minimum (Min) values, respectively, after excluding outliers.

In this study, using boxplots, outliers are estimated to be those data values that are sparsely extreme outside the expected range of variation of the data (the whiskers). These values are plotted as dots and fall higher or lower outside the end of the whiskers as seen in **Figure 4.6**. Outliers can be statistically estimated from boxplot using the interquartile range as shown in equations 4.25, 4.26, and 4.27, where IQR is the interquartile range, $Q3$ is the third quartile, and $Q1$ is the first quartile.

$$IQR = Q3 - Q1 \tag{4.25}$$

$$higher\ outliers\ values = Q3 + 1.5 * IQR \tag{4.26}$$

$$lower\ outliers\ values = Q1 - 1.5 * IQR \tag{4.27}$$

It is worth noting that areas with clusters of high or low data points were not considered as outliers (seen in) because these clusters are mostly transitional zones where each lithological class transitions gradually into another. Where outliers were detected ( ), the outliers were replaced by values estimated by linear interpolation method (equation **4.26**) using recorded known data values above and below the outlier point. **Table 4.5** and **Table 4.6** lists the number of data points identified as outliers in logging feature for each lithology in boreholes A and B, respectively. To view the boxplots of all the lithologies in both borehole A and B click the supplementary file link https://github.com/obehi-atita/Lithology-Classification-using-ML

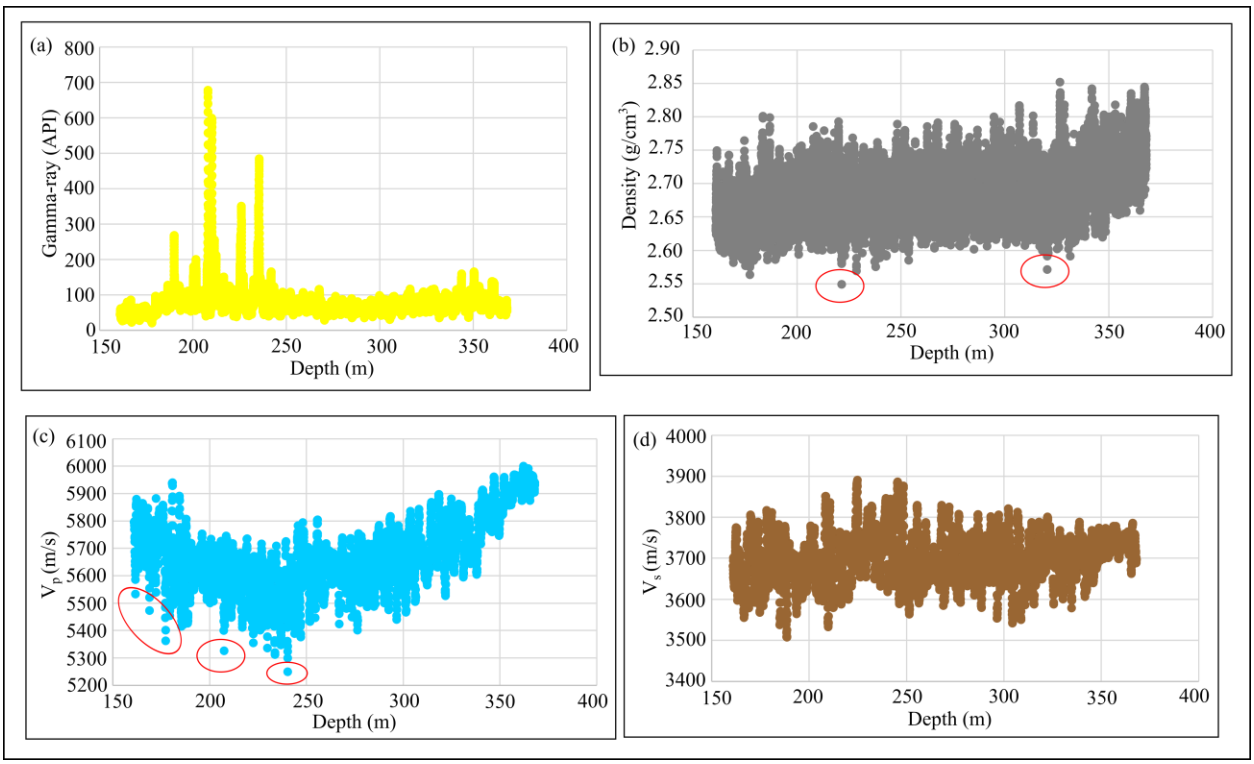

Figure 4.5: Demonstration of the relationship between logging features and depth for Upper Roodepoort quartzite in borehole A for the identification of outliers. (a) Gamma ray log, (b) Short spaced density log, (c) V_p log, and (d) V_s log.

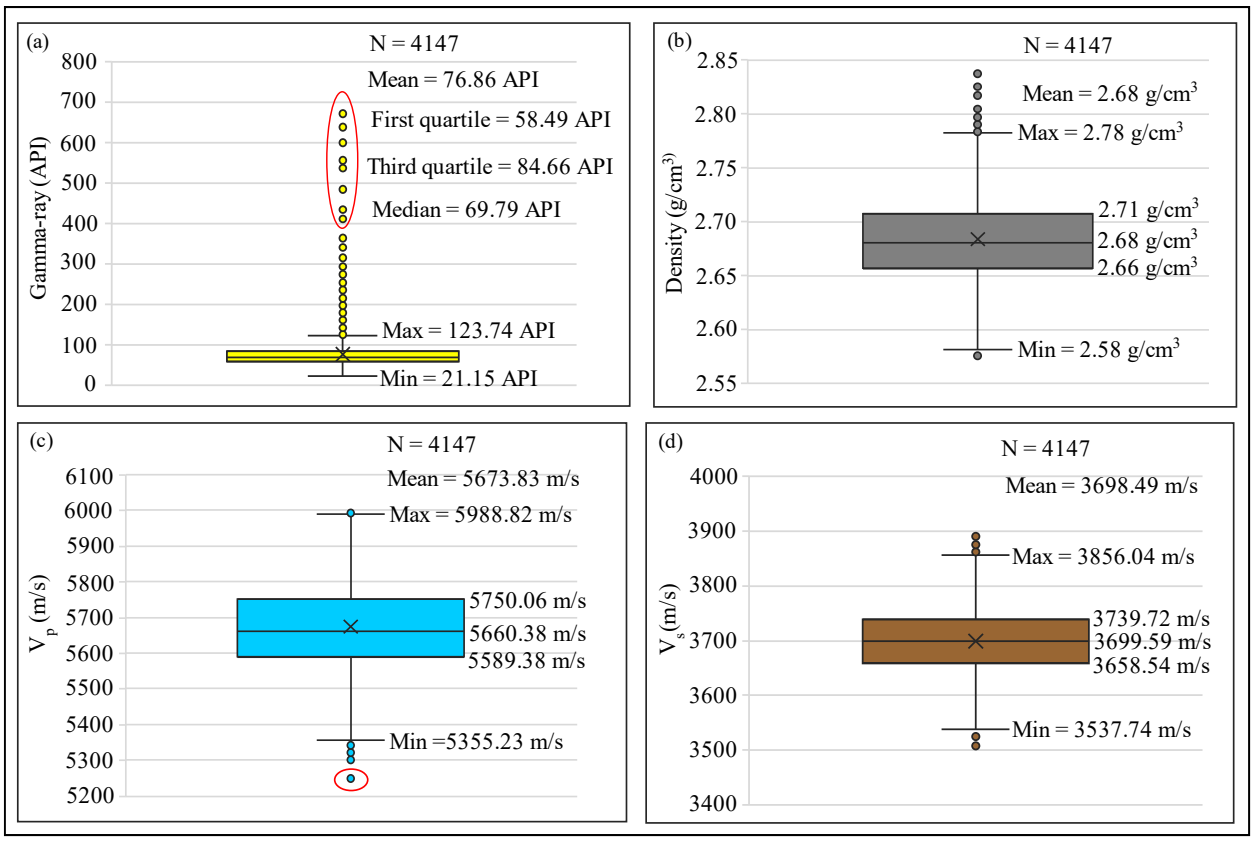

Figure 4.6: Boxplots of logging features for Upper Roodepoort quartzite in borehole A with identified outliers circled in colour red: (a) Gamma ray log, (b) Short spaced density log, (c) V_p log, and (d) V_s log. N denotes dataset size. The center line in the coloured box demonstrates the median for the data. The lower and upper boundaries of the coloured box show the first and third quartile, respectively. Lines that extend from the box are whiskers, which show the variation of each log data. The ends of the whiskers represents the maximum (Max) and minimum (Min) values in the log dataset after excluding outliers.

Table 4.5: Number of data point identified as outliers for each lithology. N denotes total number of data points; UR denotes Upper Roodepoort Formation; and LR Lower Roodepoort Formation.

Logging features	UR quartzite N = 4147	LR siltstone N = 1107	LR diabase intrusion N = 942	LR quartzite N = 711	Crown basalt N = 501	Crown diabase intrusion N = 783	Babrosco quartzite N = 2923	Babrosco diabase intrusion N = 1323
Gamma ray	8	5	5	2	0	0	0	0
density	0	0	0	7	1	0	0	0
V_p	1	2	0	3	12	10	2	2
V_s	0	1	0	0	6	5	2	2

Table 4.6: Number of data point identified as outliers for each lithology in borehole B. N denotes total number of data points; UR denotes Upper Roodepoort Formation; and LR Lower Roodepoort Formation.

Logging features	UR siltstone N = 1012	UR diabase intrusion N = 300	UR quartzite N = 4721	LR siltstone N = 1059	LR diabase intrusion N = 741	LR quartzite N = 1300	Crown basalt N = 200	Crown diabase intrusion N = 1040
Gamma ray	7	0	5	3	4	0	0	0
density	1	0	0	3	4	0	9	3
V_s	2	0	5	7	1	1	2	2

The final plot of the preprocessed (missing data and outliers treated) data is shown in **Figure 4.7**, with a total of 13091 and 10373 data points in Hole A and B, respectively. **Figure 4.7** shows data with no missing data values and less variability, which is suitable for the implementation of proposed machine learning algorithms. However, the data size of log features for lithological class is still imbalanced (a differential on the order of data point magnitude for some classes of lithologies compared to other classes in the dataset), which may cause performance reductions in prediction as the machine learning algorithms tend to be bias towards the prediction of a lithology class with larger data size. Also, lithology misclassification is expected to occur at the boundary zones with clusters of high- and low- data points.

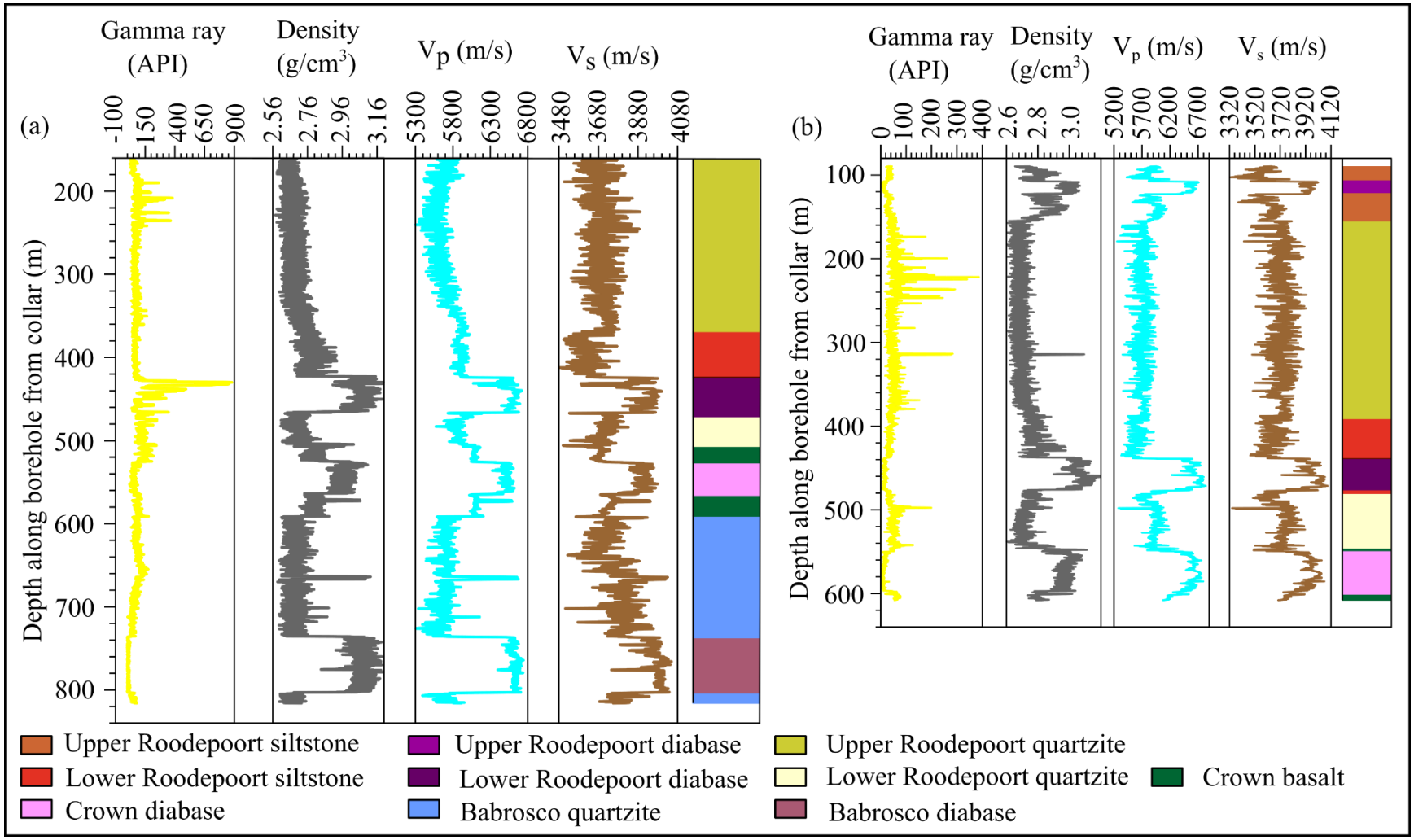

Figure 4.7: Preprocessed geophysical logs of: (a) Borehole A, and (b) Borehole B. Profile sections are colour-coded based on lithology.

4.1.3.3. Dataset Standardization

Standardization (also known as normalization) of dataset is a basic data preprocessing step in machine learning classification that transforms data points into a dimensionless smaller interval without changing the sequence of the data points. Each type of logging dataset often has different dimensions, scaling units, and significant difference in numerical order of magnitudes, which may cause the classification performance outcomes of the machine learning-based algorithms to be controlled by those data with larger magnitudes. For example, K-means clustering algorithm that implements the Euclidean distance to construct clusters may be biased on logging features with relatively larger numerical magnitudes. Standardizing the original logging feature dataset is necessary to remove prediction bias caused by difference in dimensions and scaling on machine learning-based algorithm's classification performance to improve predictions accuracy. All the preprocessed data (**Figure 4.7**) were standardized using the mean and standard deviation of feature dataset on each data point as shown in equation **4.28** (Sun et al., 2019). After standardization, logging feature dataset become dimensionless and normally distributed with similar order of magnitude such that the mean of all of the values of each logging dataset is 0 and the standard deviation is 1 (see **Table 4.7**Error! Reference source not found.), which is satisfactory for in-depth and comparative analysis. Standardization also supports data reuse and reduces the effects of outliers on machine learning-based algorithms (Zhang et al., 2022).

$$std(x_i) = \frac{x_i - \mu_x}{\sigma_x} \tag{4.28}$$

where $std(x_i)$ denotes standardized value for the ith data point, x_i denotes actual recorded data value for the ith data point, μ_x denotes mean value of the sample dataset and σ_x denotes standard deviation value of the sample dataset.

Table 4.7: Sample snippet of standardized dataset in borehole A.

Gamma ray	Density	V_p	V_s
-0.12	1.2	1.5	1.2
-0.24	-0.60	-0.94	0.00
-0.01	-0.63	-0.80	-1.23
…	…	..	..
0.3	-0.42	-0.44	-0.01
-0.34	0.31	-0.54	0.27
-0.14	-0.66	-0.62	-0.78

4.2. Supervised Lithostratigraphic Classification

4.2.1. Data Splitting

The preprocessed log dataset from boreholes A and B (**Figure 4.7**) were randomly split automatedly using a Python script (see code in section **4.1.2**) into non-overlapping training and testing datasets consisting of 80% and 20% of the total datasets, respectively. The training: testing split of 80:20 was implemented because it displayed the highest prediction accuracy for all the four models when compared with other data splitting ratios (**Figure 4.8**). The training and testing datasets have a total of 10467 and 2617 datapoints, respectively, in borehole A; and a total of 8298 and 2075 data points of training and testing dataset, respectively, in borehole B. The training set can be considered as the part of the dataset that is used for model training and cross-validation. The testing set can be considered as unseen dataset that is reserved for assessing model performance.

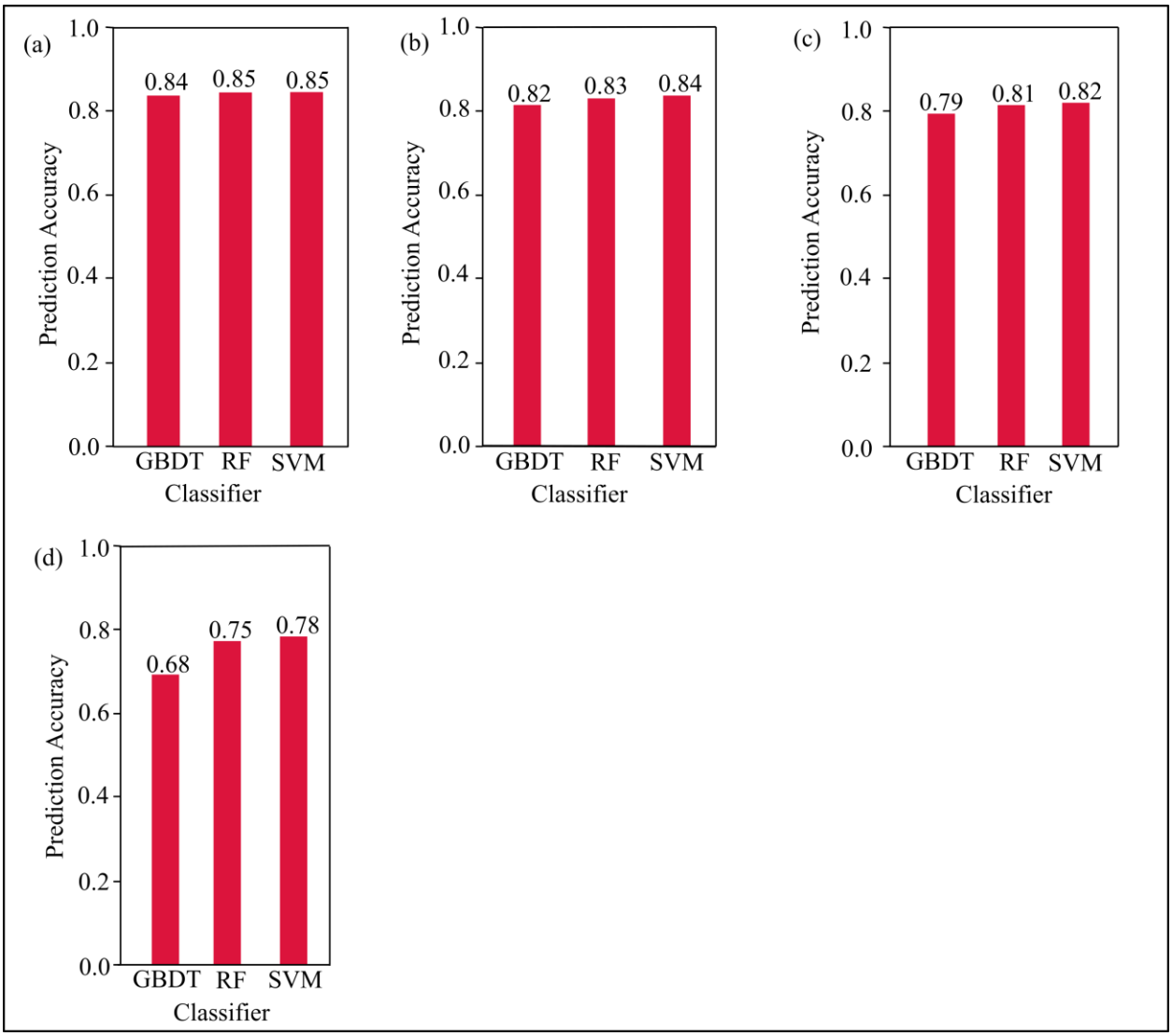

Figure 4.8: Comparison of prediction accuracy for: (a) training set: testing set = 80:20, (b) training set: testing set = 70:30, (c) training set: testing set = 60:40, and (d) training set: testing set = 50:50. GBDT, RF, and SVM denote gradient boosting decision trees, random forest, and support vector machines models, respectively.

4.2.2. Optimization and Hyperparameter Tuning

When implementing the above discussed supervised machine learning models for lithology classification problems in section **4.1.1**, each model hyperparameter and default value can impact prediction performances by either causing the model to overfit (values of parameter where the training and cross-validation scores are too far apart, e.g., a difference > 20%, causing the

generalization ability of the classifier model to be low) or underfit (when the difference between the training score and cross validation score is small, say a difference < 5%, causing the model to be unable to accurately define the relationship between the input and output features, resulting in a high error rate on both the training and testing dataset) (Xie et al., 2018; Zou et al., 2021). A selection process where a performance metric (in most cases accuracy) is applied to optimize a hyperparameter for each classifier model is known as hyperparameter tuning. The hyperparameters used and tuned for each classification model are described in **Table 4.1.** The hyperparameter tuning stage of each classifier model uses the training dataset through the grid search optimization method to automatically select the most appropriate hyperparameters combination ranges of values from which the optimal parameter value that can build a more reliable classifier model for better generalizability when tested with the unseen dataset is estimated using the 10-fold cross-validation method. Also, the validation score curve from the 10-fold cross-validation method was used with the grid search optimization method to minimize the randomness caused by different selected training data utilized for hyperparameter tuning and trade off bias for reduced variance to avoid overfitting or underfitting of each model for some hyperparameter values, which in practice enhances the classification performance (Rodriguez et al., 2009; Fushiki, 2011; Zhang et al., 2015; Konate et al. 2015; Zhao et al., 2019; Sun et al., 2019; Zou et al., 2021).

Using the 10-fold cross-validation method, the training dataset is randomly split into ten non-overlapping subsets automatically. Nine subsets are used to train the models and tune their hyperparameters, then the remaining subset is used as the validation subset to profile the prediction performances of the models. For each subset to be exploited as the validation subset, this process is repeated ten times. Then, the ten performance accuracies from the process over the validation subsets are averaged to obtain the classification accuracy for each parameter value. Finally, the parameter value that yields the highest accuracy is estimated to be the optimal value that best fits the model with least degree of variance, selection bias and overfitting or underfitting of the models (Al-Anazi and Gated, 2010).

The summary of the optimization and hyperparameter tuning processes results that were used to obtain the appropriate search range and the optimal values for each parameter of evaluated classifiers are listed in **Table 4.8** for borehole A and **Table 4.9** for borehole B.

Table 4.8: Model performance based on hyperparameter tuning and optimal parameter value using F1 score metric in borehole A. RF, SVM, and GBDT denotes Random Forest, Support Vector Machines, and Gradient Boosting Decision Trees, respectively.

Classifier model	Tuned hyperparameters	Search parameter grid	Optimal parameter value	Training time (hours)	Training F1 score	Testing F1 score
RF	number of estimators	10 , 20, 30, 40, 50, 60, 70, 80, 90, 100	100	13.92	0.9492	0.8468
	maximum depth	6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18	14			
	minimum samples split	4, 6, 8, 10, 12, 14, 16, 18, 20, 22, 24, 26, 28, 30	6			
	maximum features	Auto or sqrt	Auto			
	Bootstrap	True or False	True			
SVM	C	10^1, 10^2, 10^3, 10^4, 10^5, 10^6	10^1	34.59	0.8907	0.8457
	Gamma	1, 2, 4, 6, 8, 10	8			
	Kernel	linear, RBF	RBF			
GBDT	learning rate	0.01, 0.02, 0.03, 0.04, 0.05, 0.06, 0.07, 0.08, 0.09, 0.1	0.1	548.67	1.0000	0.8379
	number of estimators	10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 110, 120	80			
	maximum depth	1, 5, 10, 15, 20, 25	25			
	minimum samples split	5, 10, 15, 20, 25, 30	5			
	minimum samples leaf	1, 5, 10, 15, 20, 25, 30, 35	5			

Table 4.9: Model performance based on hyperparameters tuning and optimal parameter value in borehole B using F1 score metric. RF, SVM, and GBDT denotes Random Forest, Support Vector Machines, and Gradient Boosting Decision Trees, respectively.

Classifier model	Tuned hyperparameters	Search parameter grid	Optimal parameter value	Training time (hours)	Training F1 score	Testing F1 score
RF	number of estimators	10 , 20, 30, 40, 50, 60, 70, 80, 90, 100	90	**5.67**	0.9654	**0.8913**
	maximum depth	10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20	13			
	minimum samples split	2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30	6			
	maximum features	Auto or sqrt	Auto			
	Bootstrap	True or False	True			
SVM	C	10^2, 10^3, 10^4, 10^5, 10^6	10^3	**15.95**	0.9152	**0.8953**
	Gamma	1, 2, 3, 4, 5, 6, 7, 8, 9, 10	1			
	kernel	linear, RBF	RBF			
GBDT	learning rate	0.01, 0.02, 0.03, 0.04, 0.05, 0.06, 0.07, 0.08, 0.09, 0.1	0.01	**251.08**	0.9487	**0.8906**
	number of estimators	80, 100, 150, 200, 250, 300, 350, 400	400			
	maximum depth	5, 10, 15, 20, 25	5			
	minimum samples split	5, 10, 15, 20, 25, 30, 35, 40	30			
	minimum samples leaf	5, 10, 15, 20, 25, 30, 35, 40	5			

4.2.3. Model Building

To build each supervised model, selected models (RF, SVM, and GBDT) were retrained using the training dataset, the estimated optimal hyperparameters and values listed in **Table 4.8** and **Table 4.9****Table 4.9**.

4.2.4. Model Prediction Performance Evaluation and Interpretation

To test the generalizability of each of the fitted trained classifier models, the reserved unseen testing dataset was used to assess the prediction performance using evaluation metrics. Evaluation metrics are scalar methods that measure the classification performance of a trained classifier when tested with the reserved unseen dataset using a single score value (Hossin and Sulaiman, 2015).

Variance, bias, and noise are the three main sources of error in classifier model prediction (Zhang et al., 2021; Nwaila et al., 2022). The variance measures the model's output variations when the

input varies. The likelihood of a model defaulting to the prediction of some class of the lithology is known as bias, for example with our study, the model may be bias towards Upper Roodepoort quartzite due to its larger data size. The unavoidable source of error in prediction is the noise, which is neither variance nor bias (Zhang et al., 2021). To aid cross comparison of prediction quality across classes of lithology and analysis of models' performances for the purpose of obtaining the optimal classifier model, evaluation metric was estimated using the results from the training and testing datasets, and confusion matrix was used to visualize the prediction performance (Faweett, 2006; Nwaila et al., 2022).

In this study, a confusion matrix was used to show how many of a model's statistically predicted classes of lithology were correct and where the model got confused if incorrect predictions are made. In a confusion matrix, the true and the predicted labels are denoted by the rows and the columns, respectively. The confusion matrix's diagonal values represent the percentage (in a normalized confusion matrix) of times the model makes a correct prediction (where the prediction label matches the true label). The value in the other cells shows instances of bias where the model mispredicts a class of lithology (Zou et al., 2021). Performance accuracy can be described based on the confusion matrix as the fraction of the number of correctly predicted samples to the total number of samples, as mathematically shown below in equation **4.29** (Zou et al., 2021):

$$\text{accuracy} = \frac{TP + TN}{TP + FP + TN + FN} \tag{4.29}$$

where TP represent true positive classified samples, TN represent true negative classified samples, FP represent false positive classified samples and FN represent false negative classified samples.

Generally, accuracy is mostly used by researchers to evaluate the prediction performance of a classifier model and select the optimal classifier. It is worth noting that accuracy may not be a good measure of the degree of classifier prediction performance, especially when the data size distribution of lithology classes is not balanced (i.e., we have class of lithology with larger data size and some with smaller data size. For example, among the quartzites classes, the Upper Roodepoort quartzite has larger data size relative to Lower Roodepoort quartzite, as shown in **It** is important to accurately classify these lithologies between layers at IDCP DSeis boreholes

situated at the Jeppestown subgroup of the West Rand Group of Witwatersrand Supergroup for safety mapping and objective decision making.

Table 2.1). This is because the accuracy metric computes the average between the return of the classes within a dataset. Consequently, it can lead to a false model classification performance against classes of lithology with relatively smaller data size and poor optimal classifier model selection. However, accuracy has advantages of being easy to compute and understand and applicable for multi-class prediction (Hossin and Sulaiman, 2015; Bressan et al., 2020).

Owing to the above listed limitations and to the fact that the data size distribution of the applied labelled lithological classes in our study is imbalanced (as seen in **It** is important to accurately classify these lithologies between layers at IDCP DSeis boreholes situated at the Jeppestown subgroup of the West Rand Group of Witwatersrand Supergroup for safety mapping and objective decision making.

Table 2.1), accuracy might not be suitable for the selection of the optimal classifier model in this study. Hence, the performance metric F1 score (Nwaila et al., 2022) was applied for the measure of the classification performance of a classifier and selection of an optimal classifier model. An F1 score is a little less intuitive metric, which harmonically combines precision and recall metrics into one metric with the best score being 1 and the worst score being 0 (Nwaila et al., 2022). Precision on the classification report is the number of accurately identified members of a class divided by the total number of predictions made for that class. It can be mathematically represented as seen in **equation 4.30** (Xie et al., 2018; Zou et al., 2021). Recall is the number of members of a class that the classifier identified accurately divided by the total number of members in that class and can be expressed mathematically as seen in **equation 4.31** (Xie et al., 2018; Zou et al., 2021).

The F1 score will be high if both precision and recall are high. F1 score will be low if one is high, and the other is low. F1 score can be mathematically represented as seen in **equation 4.32** (Xie et al., 2018; Zou et al., 2021).

$$\text{precision} = \frac{TP}{TP + FP} \tag{4.30}$$

where TP represent true positive classified samples and FP represent false positive classified samples.

$$\text{recall} = \frac{TP}{TP + FN} \tag{4.31}$$

where TP true positive classified samples and FN false negative classified samples.

$$\text{F1 score} = \frac{2 * precision * recall}{precision + recall} \tag{4.32}$$

4.3. Impact of Linear Correlation Between Logging Features on Classification Performance

Using the same workflow (**Figure 4.4**), the effect of the linear correlation between logging features in borehole A on the classification performance of our classifiers (RF, SVM and GBDT) can be evaluated. Firstly, the degree of linear correlation between the logging features was analyzed. Then, Principal Components Analysis (PCA) was conducted on the features to rotationally transform and uncorrelate the datasets. The retained principal component feature(s) is used as input and output features to train the same applied models and evaluate the prediction performance of each model. Finally, the prediction performance based on the PCA transformed data is compared with the prediction performance based on the original preprocessed logging data. The effect of linear correlation between logging features on classification performance of the classifiers in borehole B will not be evaluated due to the poor quality log data.

4.3.1. Analysis of The Correlations Between Log Features

The geophysical logging features are usually linearly correlated. A phenomenon called parameter redundancy may arise when logging features are significantly correlated, or the sample size is large. This phenomenon may impact the classification performance and the computation time of machine learning models. Hence, correlation analysis must be performed to determine the degree of correlation among features for the acquired logging dataset (Sun et al., 2019) in order to determine the suitable method to uncorrelate. Linear correlation analysis method was used in this study to assess the correlation among the log features.

4.3.2. Principal Components Analysis (PCA)

Principal components analysis is a kind of data exploration analysis that linearly reduces the number of dimensions of a dataset. PCA uses the redundancies present in a multivariate dataset of probably correlated variables to create a smaller uncorrelated dataset variables called Principal Components (PCs) retaining almost the same amount of information as the original dataset (Benaouda et al., 1999; Kazmierczuk and Jarzyna, 2006; Rencher, 2005; Konate et al. 2015). So, one PC can be analyzed without referring to others. PCA has been successfully applied to logging data to resolve a wide range of problems in geophysics (Rezaie et al., 2012; Xu et al., 2006; Gelfort, 2006; Konate et al. 2015). PCA is also conducted to make further analysis and visualization of dataset using other techniques (e.g., clustering, classification) easier. PCA is based on distance calculations to define the similarity between datasets. PCA involves the statistical estimate of the eigenvalues and cumulative variance of a preprocessed dataset (see data preprocessing section 4.1.3), and the principal component(s) with most information are retained.

For the selection of the optimal principal component for interpretation for this study, the Kaiser (1960) criterion, which proposes that principal component(s) with eigenvalues greater than 1.0 be retained for further analysis and interpretation, was adopted. However, the first principal component will probably contain the maximum information from the original dataset, followed by the second principal component and so on until the last principal components, which contains the least amount of information (Benaouda et al., 1999).

PCA has noticeable drawback: it is often challenging to interpret the resulting PC(s) because all of the original n variables are combined linearly to each PC, and in most cases the coefficients of the linear combination (loadings, i.e., how much each original variable contributes to the corresponding PC) are non-zero. Hence, rotation techniques are mostly applied to guide researchers in interpreting PC (Rencher, 2005; Konate et al. 2015).

4.4. Unsupervised Lithostratigraphic Classification

For K-means clustering analysis, the number of clusters (K) the datapoints are grouped into is a hyperparameter. The optimal value of K was determined in this study using the 'Elbow' method. The 'Elbow' method involves running a K-means model on the datasets for a range number of K putting in consideration the number of classes of lithology from the core logging. In this study, the number of K ranging between 1, 2, 3, 4, 5, 6, 7, 8, 9, and 10 was used to optimize the K-means model. For each random value of K, the sum of squared distance is calculated. The sum of squared distance is calculated as the average of the squared distances from the cluster centers of each cluster. Then, a plot of the calculated sum of squared distance against the different random values of K is generated. Accordingly, the value of optimal K is selected to be that value at the 'Elbow point', that is at a random value of K after which the sum of squared distance begins to experience a linear decrease. To ensure the algorithm converges to global minimum, centroids initializations with minimum sum of squared distance are used (Lloyd, 1982; Forgy, 1965; Benaouda et al., 1999; Kazmierczuk and Jarzyna, 2006). The K-means model was built using the estimated optimal number of K using the defined workflow in **Figure 4.4**. K-means clustering was only performed on data from borehole A due to the fact that the log features in borehole B are highly overlapped, which makes it difficult to perform K-means clustering.

5. RESULTS

5.1. Supervised Lithostratigraphic Classification in Borehole A

5.1.1. Optimization and Hyperparameter Tuning

This section of chapter describes results of the optimization and hyperparameter tuning processes that were used to obtain the appropriate search range and the optimal values for each parameter of evaluated models (RF, SVM and GBDT) using grid search and 10-fold cross-validation methods (as discussed in section **4.2.2**).

Figure 5.1 and **Figure 5.2** illustrates the results of validation curve and 10-fold cross-validation application to the Random Forest (RF) classifier model using the selected hyperparameters and computed optimal values for each parameter listed in **Table 4.1**. **Figure 5.1** shows that for optimal training of RF to avoid overfitting (values of parameter where the training and cross validation scores are too far apart, say a difference > 20%, causing the generalization ability of the classifier

model to be low) using the grid search optimization method, the applicable combined parameter search range for number of trees (n_estimators) is from 10 to 100 trees in the increment of 10, maximum depth (max_depth) is from 6 to 18 in the increment of 2 and minimum samples split (min_samples_split) is from 4 to 30 in the increment of 2. RF hyperparameters tuning using the search parameter grids in **Table 4.1** and **Figure 5.1**, the 10-fold cross-validation method yields optimal value of 100, 14, 6, and 0.8503 for number of trees, maximum depth, minimum samples split and performance accuracy, respectively (**Figure 5.2**).

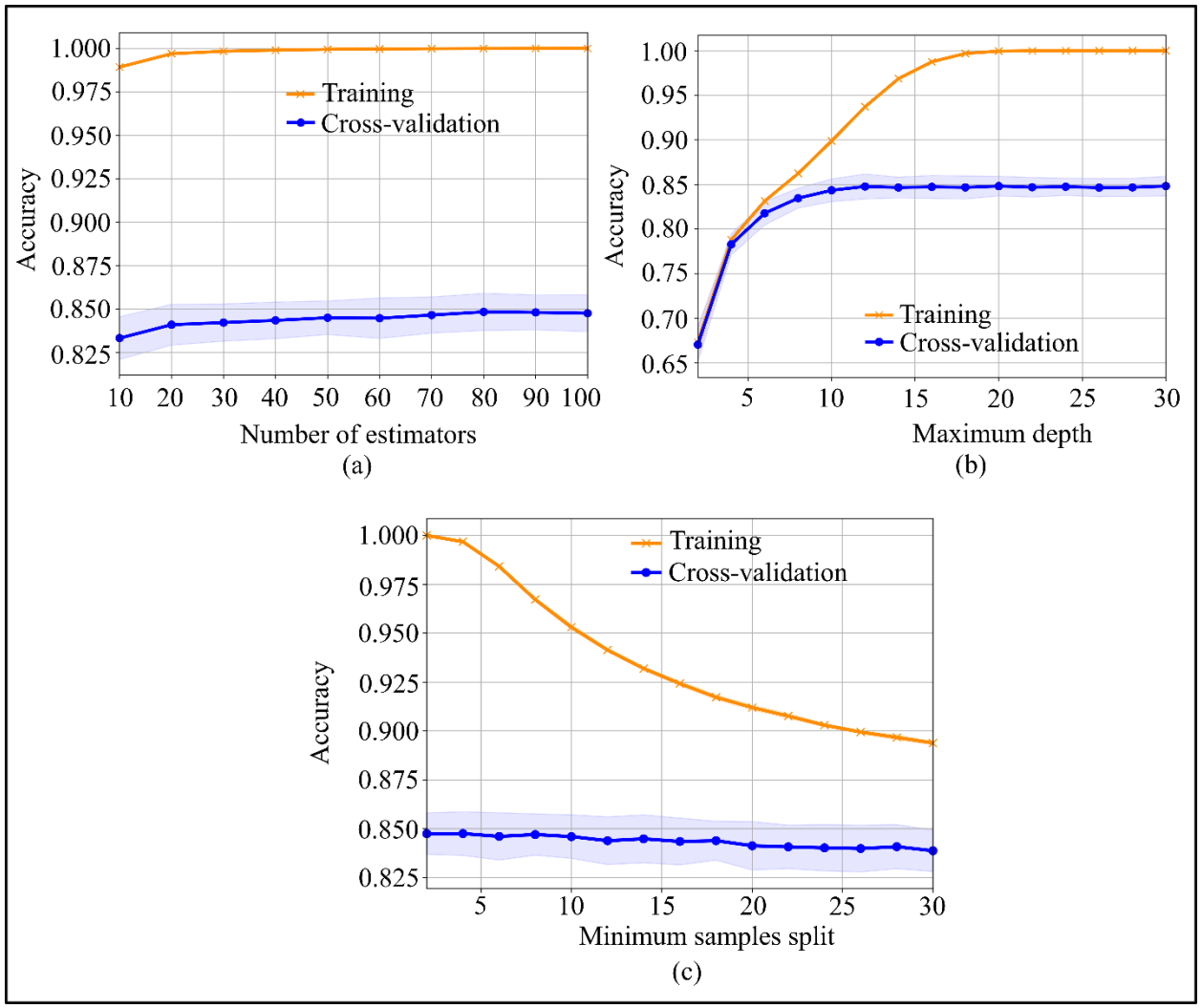

Figure 5.1: Validation curve for Random Forest classifier's hyperparameters using grid search method in borehole A: (a) required trees to construct the forest, (b) individual forest maximum growth depth, and (c) required minimum samples to split an internal node.

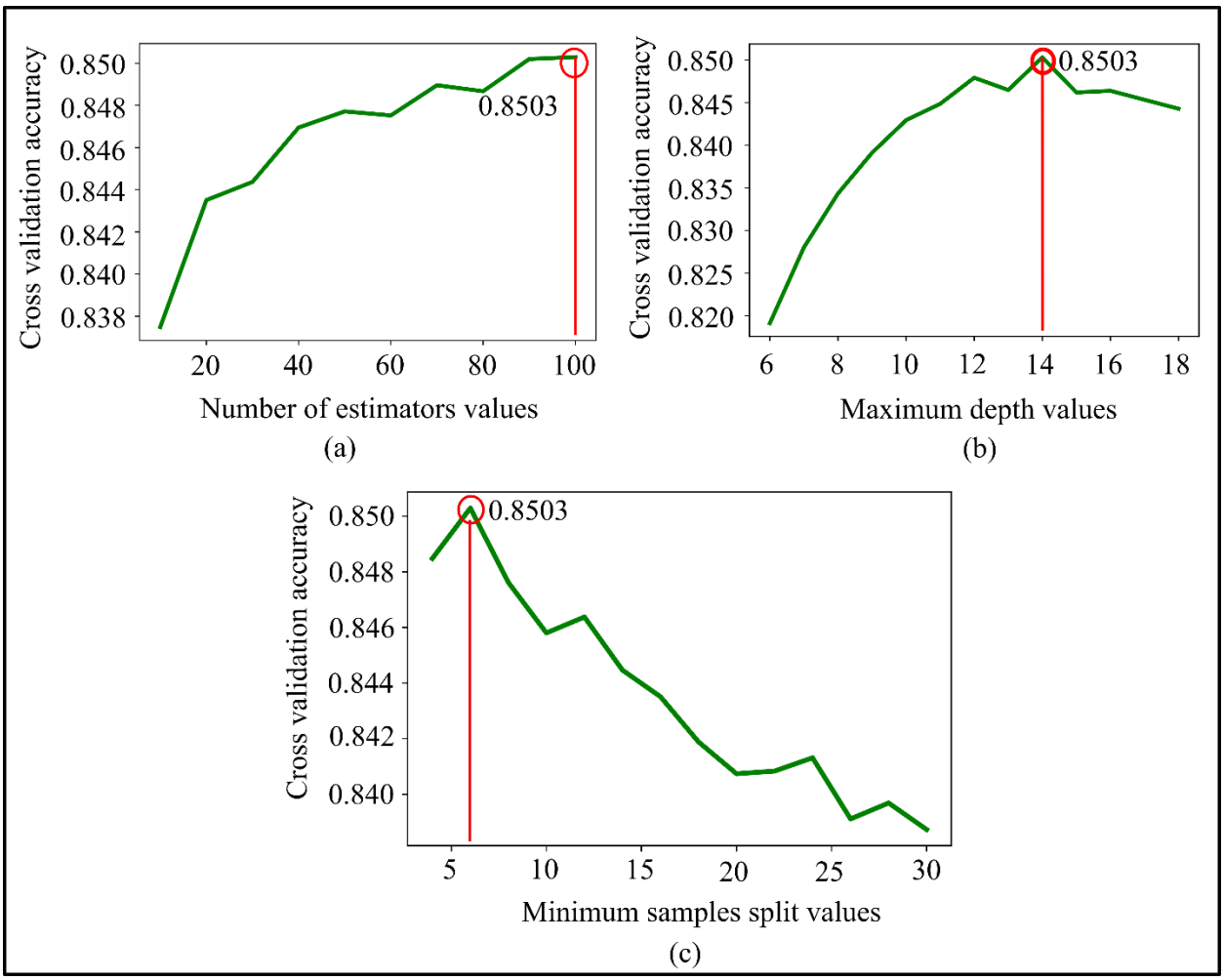

Figure 5.2: 10-fold cross-validation of the random forest model's tuned hyperparameters to yield optimal accuracy of 0.8503 in Borehole A: (a) Optimal value of number of trees when optimal values of maximum depth =14 and minimum samples split = 6, (b) Optimal value of maximum growth depth when optimal values of number of trees = 100 and minimum samples split = 6, and (c) Optimal value of minimum samples split when optimal values of number of trees = 100 and maximum depth = 14.

Using search parameter ranges [10, 100, 1000, 1000, 10000, 100000, 100000] for penalty constant parameter (C), [1, 2, 4, 6, 8,10] for the kernel spread parameter (Gamma) and RBF kernel as appropriate parameter search ranges to optimize SVM model based on the validation curve (**Figure 5.3**), the optimal values of parameter C and Gamma were obtained to be 10 and 8, respectively, via the grid search and 10-fold cross-validation methods. **Figure 5.4** shows that the combination

of these optimal values of parameters C and Gamma of the SVM model yielded the highest performance accuracy of 0.8495.

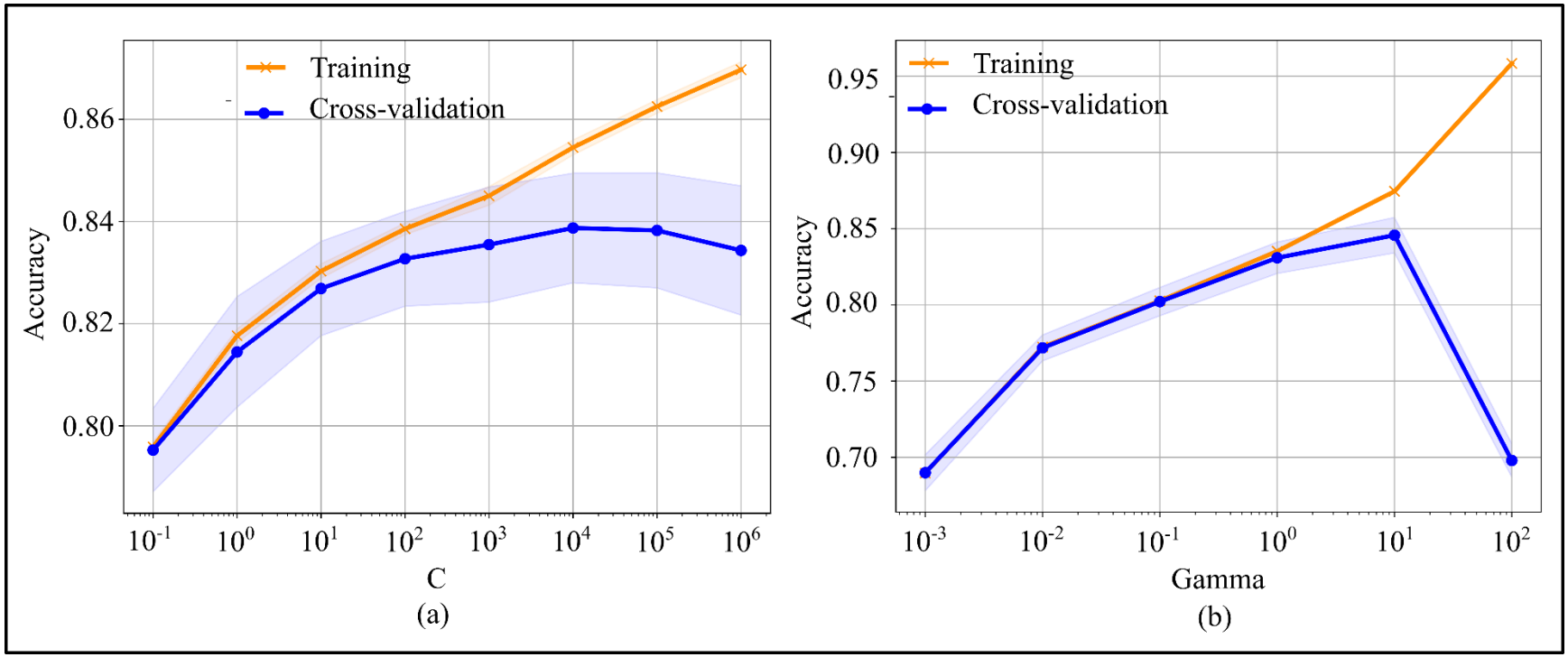

Figure 5.3: Validation curve for the Support Vector Machine classifier's hyperparameters using grid search in borehole A: (a) Penalty constant parameter (C), and (b) Kernel spread parameter (Gamma).

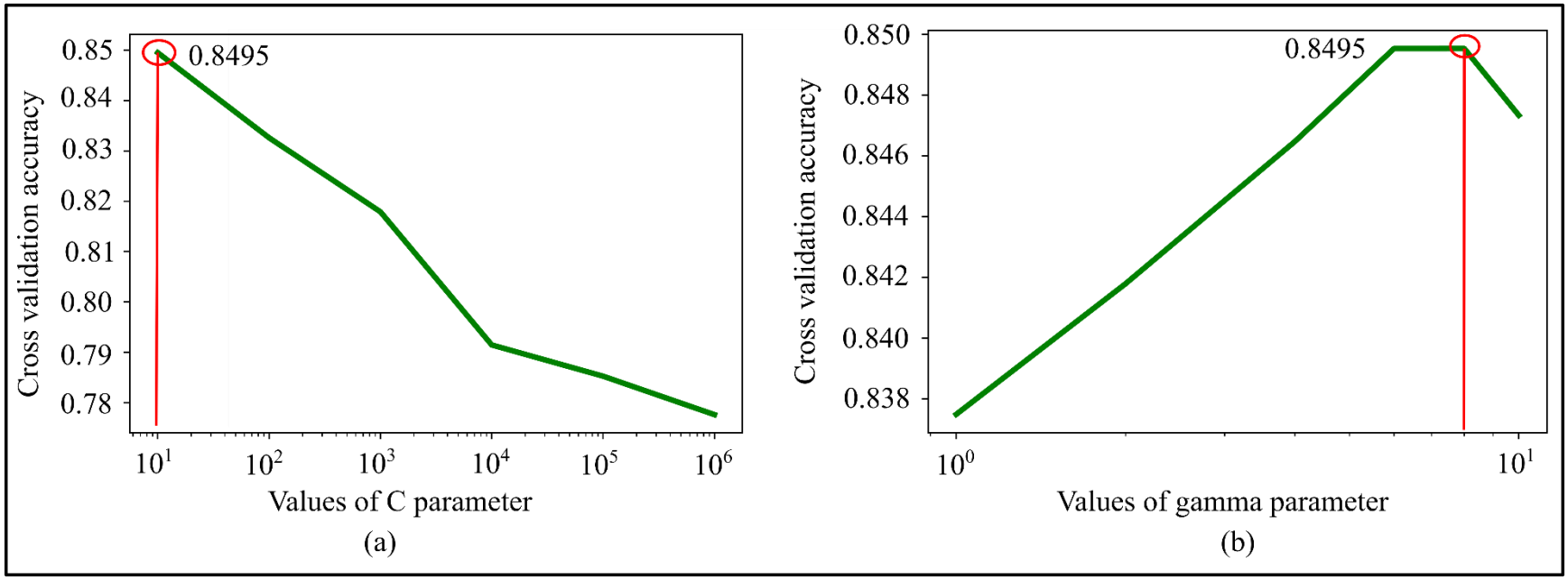

Figure 5.4: 10-fold cross-validation of the Support Vector Machine model's tuned hyperparameters to yield optimal accuracy of 0.8495 in borehole A: (a) Penalty constant parameter (C)when the optimal value is 10, and (b) Kernel spread parameter Gamma when the optimal value is 8.

For optimal prediction performance of GBDT, the results of grid search method in **Figure 5.5** demonstrate that the proper parameter grid required for number of trees is 10 to 120 in the increment of 10, learning rate is 0.01 to 0.1 in the increment of 0.01, minimum samples leaf is [1, 5, 10, 15, 20, 25, 35], minimum samples split is 5 to 30 in the increment of 5, and maximum depth is [1, 5, 10, 15, 20, 25]. **Figure 5.6a** illustrates that GBDT optimal parameter value of number of trees is 80 when that of maximum depth = 5, minimum samples leaf = 25, learning rate = 0.1, and minimum samples split = 10. **Figure 5.6b** shows the optimal value of the parameter maximum depth is 5 when number of trees = 80, minimum samples leaf = 25, learning rate = 0.1, and minimum samples split = 10. **Figure 5.6c** illustrates that the optimal parameter value of min samples leaf is 25 when that of number of trees = 80, min samples split = 10, learning rate = 0.1, and maximum depth = 5. **Figure 5.6d** shows the optimal parameter value of minimum samples split is 10 when that of number of trees = 80, maximum depth = 5, learning rate = 0.1, and minimum samples leaf = 25. **Figure 5.6e** shows that the optimal parameter value of learning rate is 0.1 when number of trees = 80; minimum samples split = 10, minimum samples leaf = 25, and maximum depth = 25. Finally, the combination of these optimal parameters' values yielded the optimal performance accuracy of 0.8461.

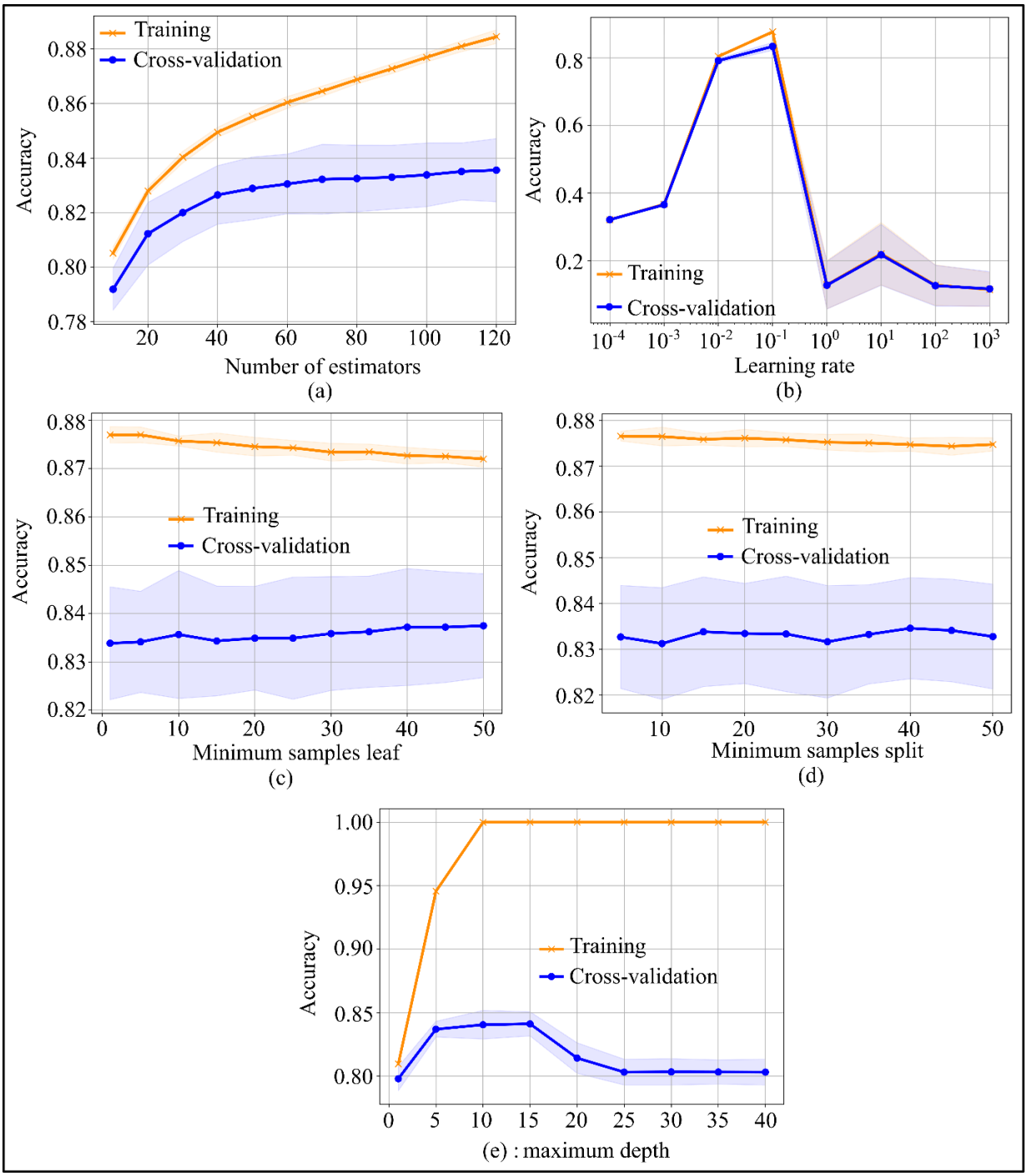

Figure 5.5: Validation curve for the GBDT classifier's hyperparameters in borehole A: (a) Required number of trees, (b) Learning rate, (c) Required minimum samples at a leaf node, (d) Required minimum samples to split an internal node, and (e) Individual tree maximum growth depth.

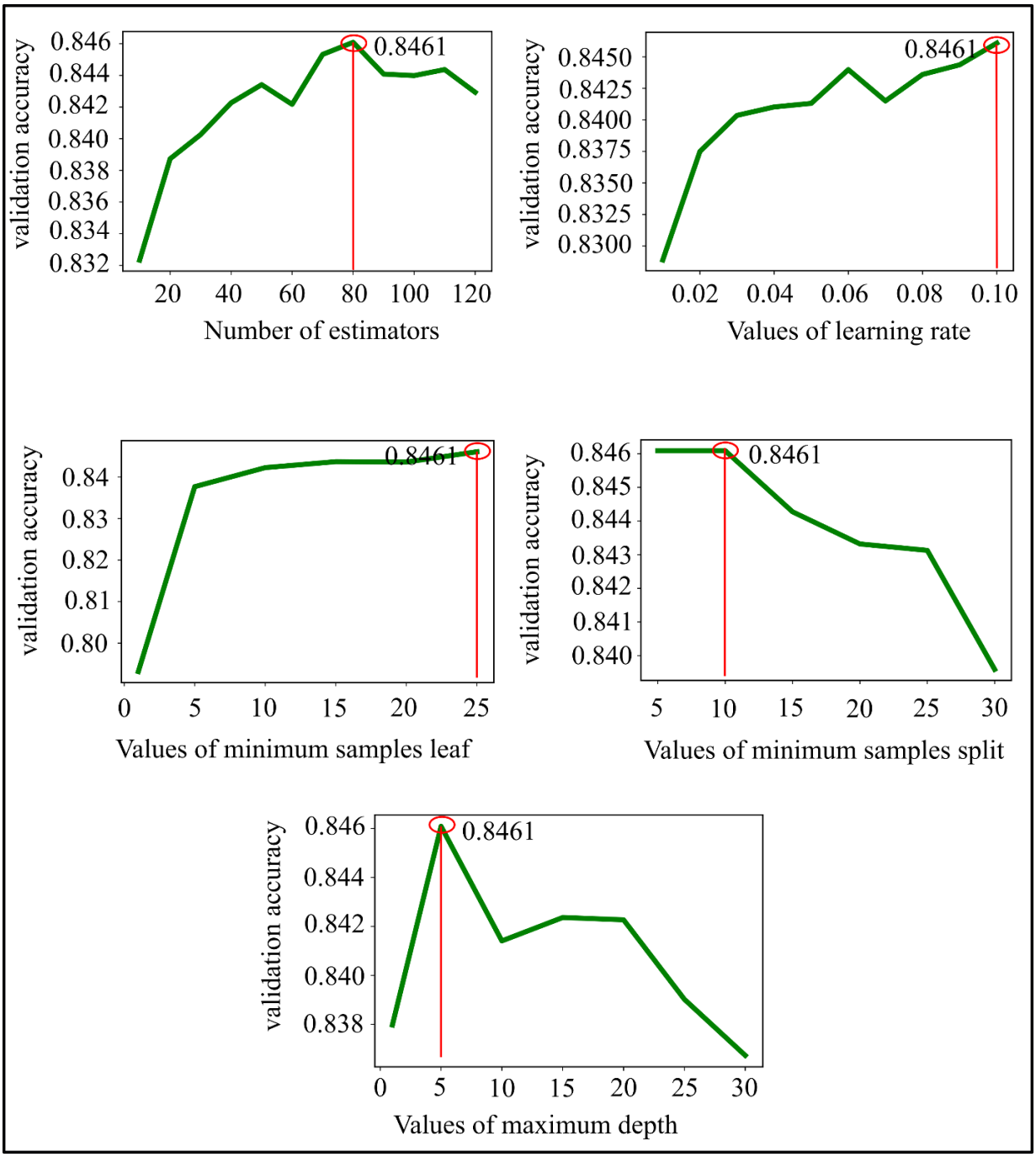

Figure 5.6: 10-fold cross-validation of the Gradient Boosting Decision Tree classifier's hyperparameters in borehole A: (a) Optimal value number of trees, (b) Optimal value of learning rate, rate, (c) Optimal value of minimum samples leaf, (d) Optimal value of minimum samples split, and (e) Optimal maximum growth depth of individual tree.

5.1.2. Performance Evaluation Metrics

To evaluate the implemented model (RF, SVM and GBDT) prediction performances for cross-comparison and selection of optimal model purposes, the F1 score, precision and recall method metrics of each lithology class based on the training and testing datasets using the 10-fold cross-validation method were estimated and listed in **Table 5.1** and **Table 5.2**, respectively. Generally, the classifier models: Random Forest, Support Vector Machines and Gradient Boosting Decision Tree performed similarly with F1 scores over 0.80 during model training and testing. However, the Random Forest model performed best with an F1 score of 0.8486 during model testing. Based on testing dataset results per lithological class, Lower Roodepoort siltstone, Lower Roodepoort diabase intrusion, Crown basalt, Crown diabase intrusion, and Babrosco diabase intrusion were more correctly predicted with an F1 score of over 0.90.

Table 5.1: Classifier model training performance based on each lithological class and several evaluation metrics using the **training datasets** in borehole A. UR and LR denote Upper Roodepoort and Lower Roodepoort formations, respectively. Average F1 score for a classifier model denoted in red colour.

	Gradient Boosting Decision Tree			Random Forest			Support Vector Machine		
Lithology	Precision	Recall	F1-score	Precision	Recall	F1-score	Precision	Recall	F1-score
UR Quartzite	1.0000	1.0000	1.0000	0.9088	0.9607	0.9340	0.8342	0.8759	0.8546
LR Siltstone	1.0000	1.0000	1.0000	0.9865	0.9843	0.9854	0.9542	0.9618	0.9580
LR Diabase intrusion	1.0000	1.0000	1.0000	0.9919	0.9646	0.9780	0.9727	0.9344	0.9531
LR Quartzite	1.0000	1.0000	1.0000	0.9559	0.9091	0.9319	0.8541	0.7675	0.8085
Crown Basalt	1.0000	1.0000	1.0000	0.9913	0.9771	0.9842	0.9885	0.9814	0.9849
Crown Diabase intrusion	1.0000	1.0000	1.0000	0.9578	0.9871	0.9722	0.9419	0.9662	0.9539
Babrosco Quartzite	1.0000	1.0000	1.0000	0.9455	0.8881	0.9159	0.8465	0.8122	0.8290
Babrosco Diabase intrusion	1.0000	1.0000	1.0000	0.9971	1.0000	0.9986	0.9952	0.9990	0.9971
Weighted average	1.0000	1.0000	**1.0000**	0.9501	0.9493	**0.9492**	0.8913	0.8909	**0.8907**

Table 5.2: Classifier model performance based on each lithological class and several evaluation metrics using the **testing datasets** in borehole A. UR and LR denote Upper Roodepoort and Lower Roodepoort formations, respectively. Average F1 score for a classifier model denoted in red colour.

	Gradient Boosting Decision Tree			Random Forest			Support Vector Machine		
Lithology	Precision	Recall	F1 score	Precision	Recall	F1 score	Precision	Recall	F1 score
UR Quartzite	0.7662	0.8130	0.7889	0.7735	0.8384	0.8046	0.7647	0.8270	0.7946
LR Siltstone	0.9252	0.9167	0.9209	0.9381	0.9120	0.9249	0.9296	0.9167	0.9231
LR Diabase intrusion	0.9253	0.8895	0.9070	0.9349	0.8729	0.9029	0.9310	0.8950	0.9127
LR Quartzite	0.6884	0.6835	0.6859	0.7246	0.7194	0.7220	0.7405	0.6978	0.7185
Crown Basalt	0.9634	0.9246	0.9436	0.9579	0.9146	0.9357	0.9541	0.9397	0.9468
Crown Diabase intrusion	0.8902	0.9506	0.9194	0.8743	0.9444	0.9080	0.9118	0.9568	0.9337
Babrosco Quartzite	0.7879	0.7385	0.7624	0.8090	0.7446	0.7755	0.8056	0.7416	0.7723
Babrosco Diabase intrusion	0.9964	0.9964	0.9964	0.9964	1.0000	0.9982	0.9964	1.0000	0.9982
Weighted average	0.8389	0.8380	**0.8379**	0.8469	0.8486	0.8468	0.8468	0.8460	**0.8457**

For all the optimized predictive models (RF, SVM and GBDT) used in lithology classification, the testing F1 scores are close to the training F1 scores (**Figure 5.7**), which establishes the high generalization ability of each model for future prediction. **Figure 5.7** and Table 4.7 indicates that for numerous parameter optimization of the three models, RF classifier model has the optimal performance in predicting the lithology classes most correctly with an average testing F1 score of 0.8468 and least computational training time of 13.92 hours. However, the three evaluated optimized classifiers demonstrated similar classification with a small difference in their F1 scores. The order of lithology prediction performance of models is RF > SVM > GBDT based on the testing performance metric results. Nevertheless, the degree of applied log features overlaps between some lithological classes and uneven data size of lithological classes (see Table 3.2) caused lithology prediction biases.

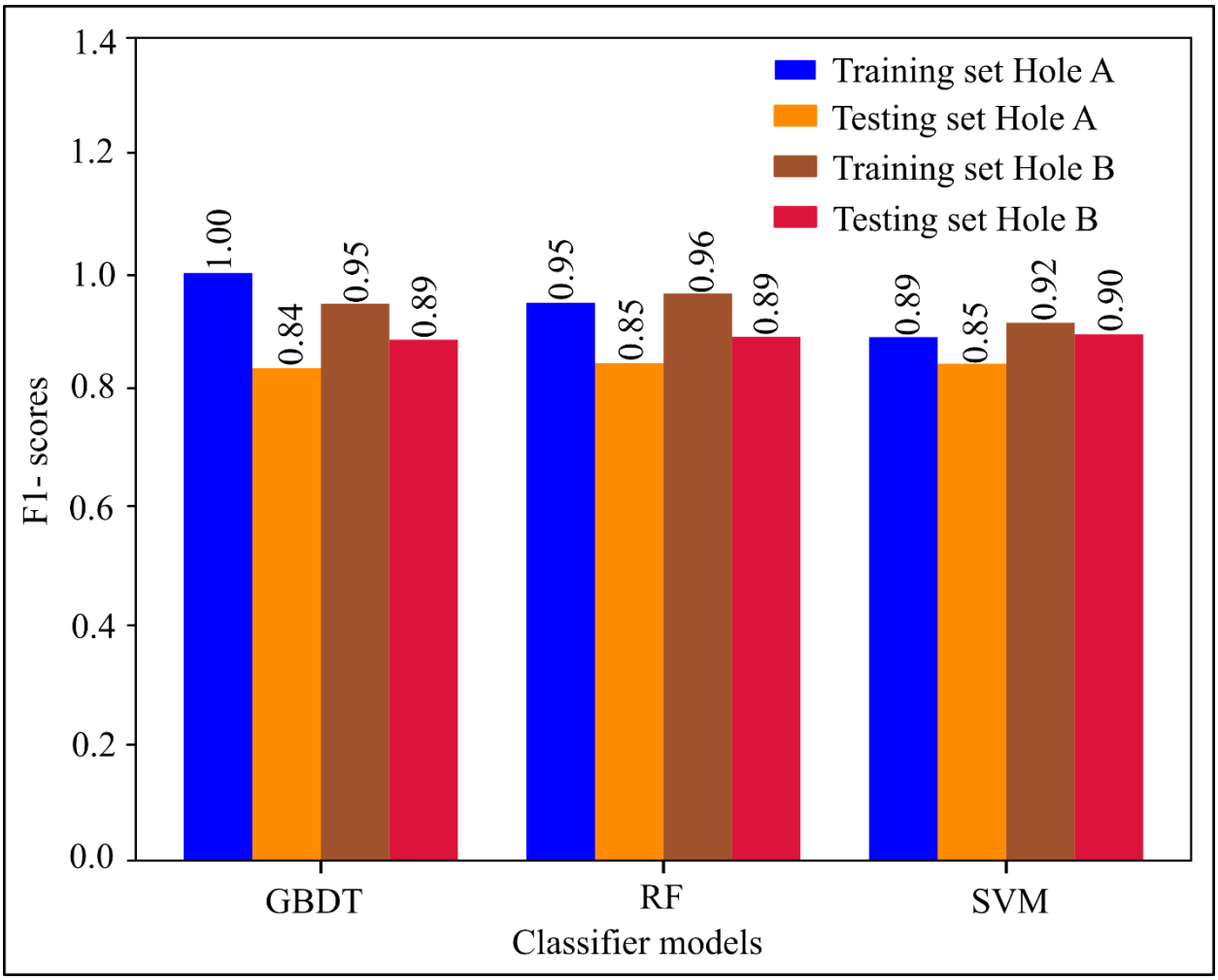

Figure 5.7: Performance comparison using the F1score metric for applied classification models (training set: testing test = 80:20 of original datasets). GBDT, RF and SVM represent Gradient Boosting Decision Trees, Random Forest and Support Vector Machines models, respectively.

5.1.3. Confusion Matrix

To visualize which classes of lithology were correctly classified or misclassified by each applied model as a means of assessing and cross comparing the classification performance of the models, confusion matrix was evaluated (**Figure 5.8**). **Figure 5.8** results show similar confusion matrices for all the models. Lower Roodepoort siltstone, Crown basalt, Crown, and Babrosco diabase intrusion appear to be easier to predict accurately with normalized confusion matrix score above

0.90. Then, to some extent (about 24%), Lower Roodepoort quartzite and Babrosco quartzite were difficult to predict.

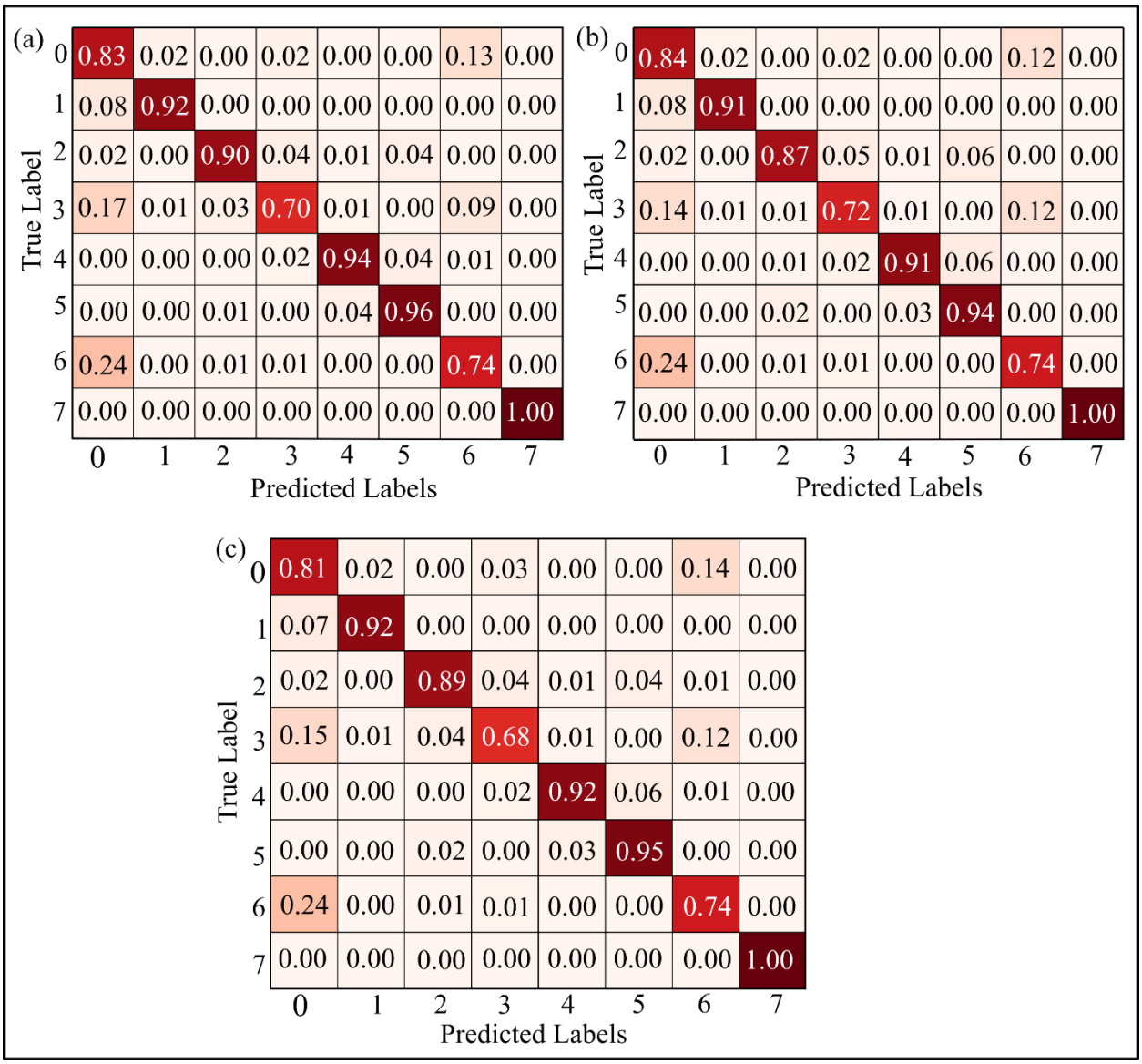

Figure 5.8: Normalized confusion matrices in borehole A based on the **testing dataset** for: (a) Support Vector Machine (SVM), (b) Random Forest (RF), and (c) Gradient Boosting Decision Tree (GBDT). The labels 0, 1, 2, 3, 4, 5, 6, and 7 denote Upper Roodepoort quartzite, Lower Roodepoort siltstone, Lower Roodepoort diabase intrusion, Lower Roodepoort quartzite, Crown basalt, Crown diabase intrusion, Babrosco quartzite, and Babrosco diabase intrusion, respectively, as encoded in **Table 4.2**.

To verify lithology classification performance of the models in borehole A, a lithological profile predicted by our machine learning models were compared with the known lithological profile from core logging conducted by Berset (2017) and Rickenbacher (2018), (**Figure 5.9**). A substantial amount of Lower Roodepoort (17% by SVM, 14% by RF, and 15% by GBDT), and Babrosco quartzite (24% by the three models) were incorrectly classified as Upper Roodepoort quartzite by the three models. Similarly, a significant amount of about 14% of Upper Roodepoort quartzite was misclassified as Babrosco quartzite by all the three models. In addition, only Random Forest (RF) and Gradient Boosting Decision Tree (GBDT) models incorrectly classified a significant proportion of about 12% of Lower Roodepoort quartzite as Babrosco quartzite. It is worth noting the core section 665.07 – 666.02 m of Babrosco quartzite that was misclassified as Lower Roodepoort diabase intrusion by RF, GBDT, and SVM models.

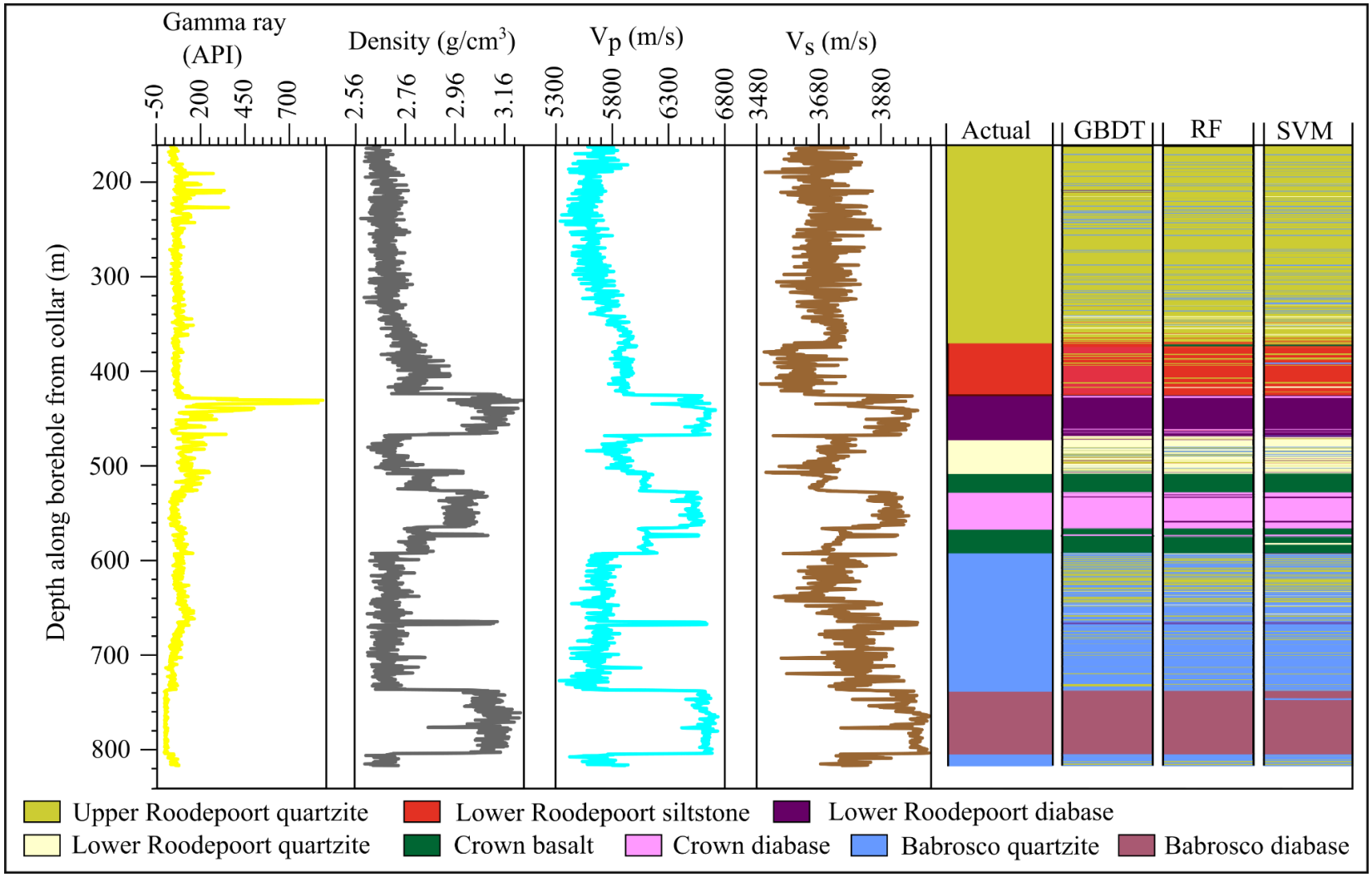

Figure 5.9: Automated lithological profile prediction using geophysical log dataset and machine learning classification models versus traditional lithological profile in borehole A. GBDT is Gradient Boosting Decision Trees, RF is Random Forest, and SVM is Support Vector Machine. Profile sections are colour-coded based on lithologies.

5.1.4. Model Interpretation

In addition to the applied performance evaluation metrics (F1 score and confusion matrix), we further elucidated the predictive strength of various logging features to the lithology classification in borehole A using feature importance ranking and Partial Dependence Plots (PDPs). Feature importance ranking profiles the predictive strength of each logging feature to the classification performance of a model, which may vary between lithologies. Partial dependence plots help comprehend interactions between the logging features and the targeted lithology classes. We evaluated feature importance through the permutation of covariates for the highest classification performing model (Random Forest). The predictive strength of logging feature to lithology classification increases with the feature importance value. For lithological classification in borehole A (**Figure 5.10** and **Table 5.3**), in the order of importance, V_p log ranks first, followed by gamma ray log, V_s log, and finally density log. This means that for classification of lithology in our study area, V_p log is the optimal logging feature needed for better accuracy.

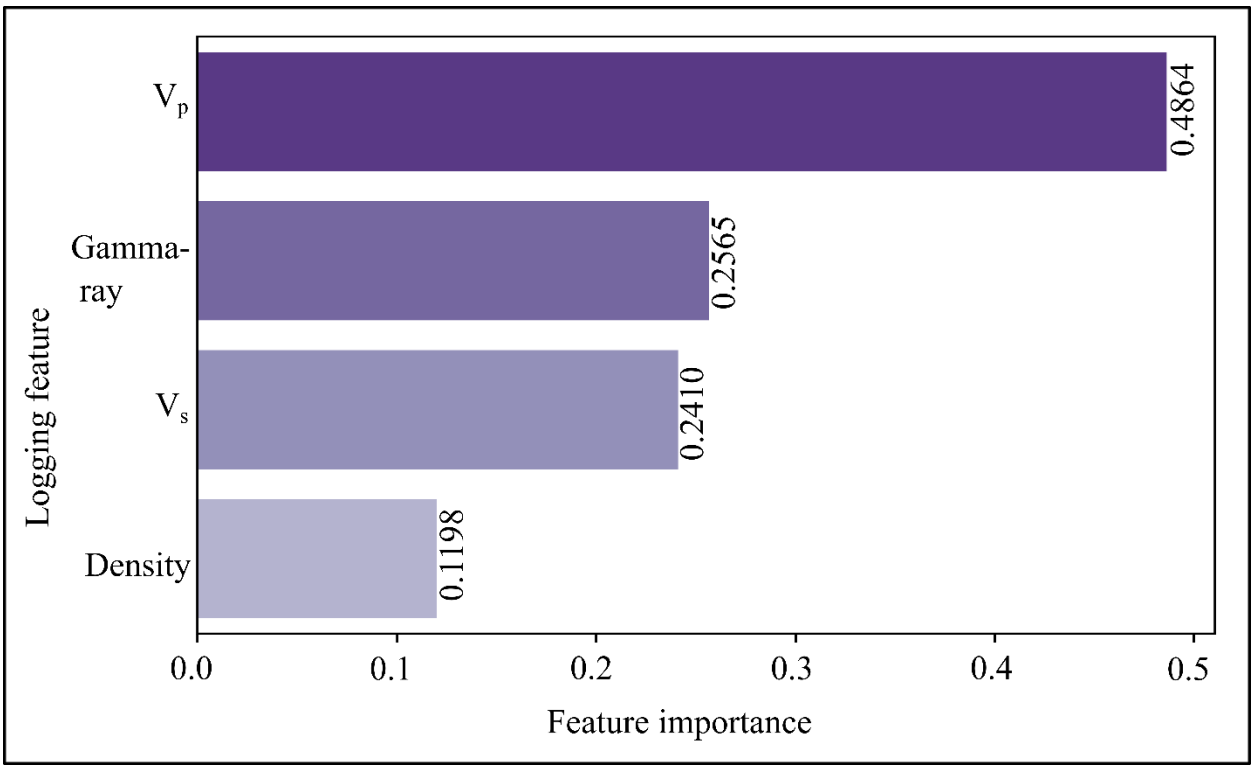

Figure 5.10: Ranking of logging features to lithology classification in borehole A by Random Forest model.

Table 5.3: logging feature variables importance to lithology classification in borehole A by Random Forest model.

Feature parameter	Feature importance
P-wave log (V_p)	0.4864
Natural gamma ray log	0.2565
S-wave log (V_s)	0.2410
Density log	0.1198

To further understand how the two most important logging features impacted (positively or negatively) on the targeted most predominate lithology class (quartzite) classification through the highest classification performing model (Random Forest) in borehole A, we used the partial dependence plots (**Figure 5.11**; **Figure 5.12**). Partial dependence plots help comprehend interactions between the logging feature and the targeted lithology classes. The followings are demonstrated by **Figure 5.11**: (1) P-wave (V_p) log value of ≤ 5400 m/s made a positive impact to the classification of Upper Roodepoort quartzite. V_p values from 5400 m/s to 6400 m/s made little impact to the lithology classification. Higher values in V_p log above 6400 m/s made no impact in the classification of lithology, (2) V_p log value of about 5400 m/s to 5600 m/s made a positive influence in the classification of Upper Roodepoort quartzite. When the value is higher from 5600 m/s to 6400 m/s, the influence of V_p log to the lithology classification increased positively, and higher values above 6400 m/s made no changes in the lithology classification, and (3) V_p logging value of about 5400 m/s to 5600 m/s made a positive impact to the classification of Babrosco quartzite. When the V_p value is higher from 5600 m/s to 5800 m/s, the impact to the lithology classification positively increased but starts to decrease when V_p value is between 5800 m/s and 6400 m/s. V_p values higher than 6400 m/s was no longer effective in the classification of Babrosco quartzite. In summary, V_p values higher than 6400 m/s will make no contributions to the classification of the quartzite classes.

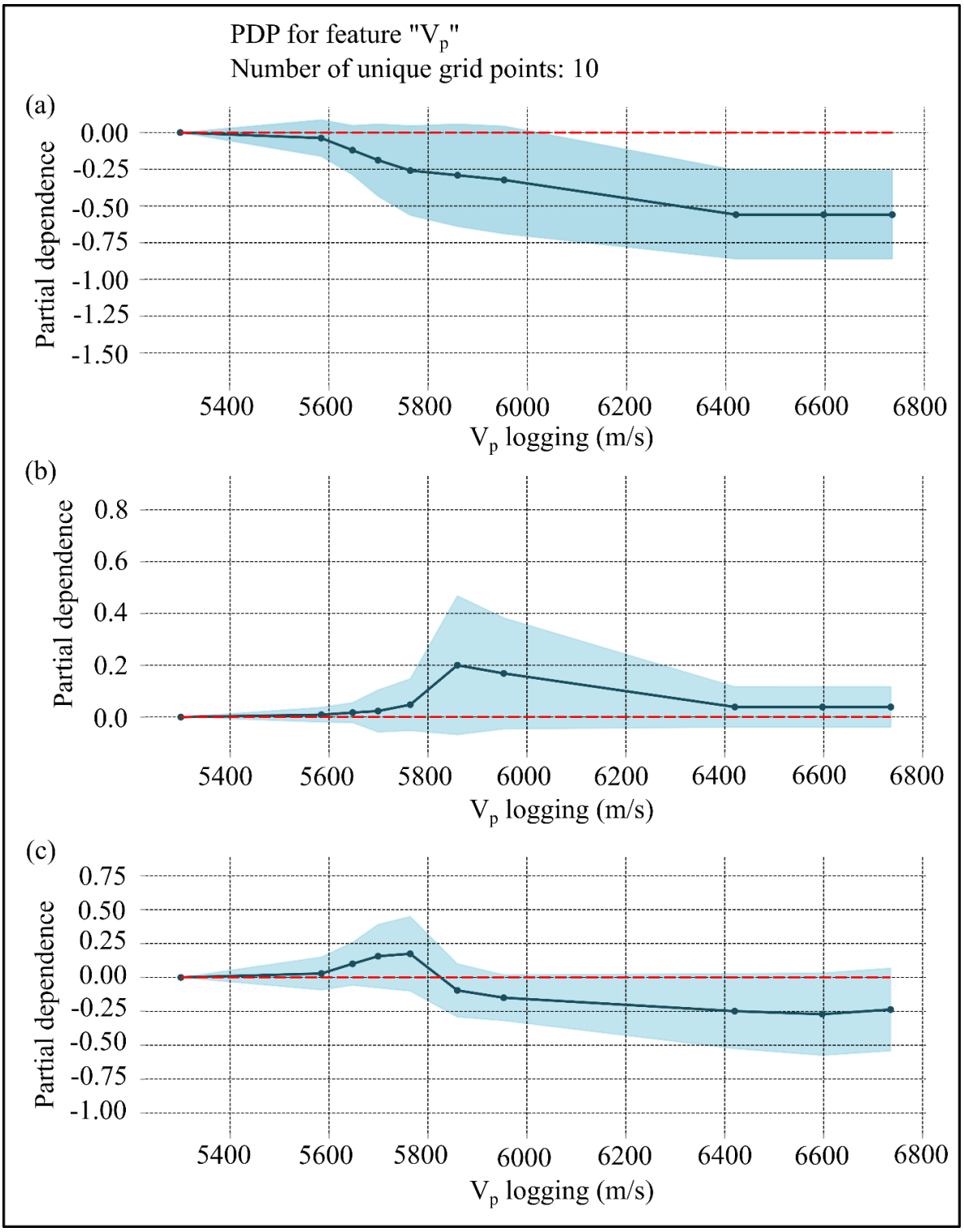

Figure 5.11: Partial dependence plots of V_p log feature in borehole A for Random Forest model classification of: (a) Upper Roodepoort quartzite, (b) Lower Roodepoort quartzite, and (c) Babrosco quartzite.

Figure 5.12 shows the following: (1) Natural gamma ray log value from 0 API to 180 API made an optimal positive effect in the classification of Upper Roodepoort quartzite. When gamma ray value is higher from 200 API to 800 API, the effect of natural gamma ray log to lithology classification decreased. When the value of natural gamma ray reached 1000 API, the contribution to lithology classification gradually increased, and when the natural gamma ray logging value is higher than 1400 API, no change is made in the contribution to the lithology classification, (2) Natural gamma ray value from 0 API to 180 API made positive effect in the classification of Lower Roodepoort quartzite. Higher value of natural gamma ray from 200 API made gradual increase in the contribution to the classification of Lower Roodepoort quartzite, and (3) Natural gamma ray value from 0 API to 1600 API made little contribution to the classification of Babrosco quartzite. In summary, the optimal value of gamma ray to the classification of Upper Roodepoort quartzite is between 0 API and 180 API. Gamma ray contributed most to the classification of Lower Rooderpoort quartzite when the value is between 400 API and 1600 API. Little contribution was made by gamma ray log to the classification of Babrosco quartzite.

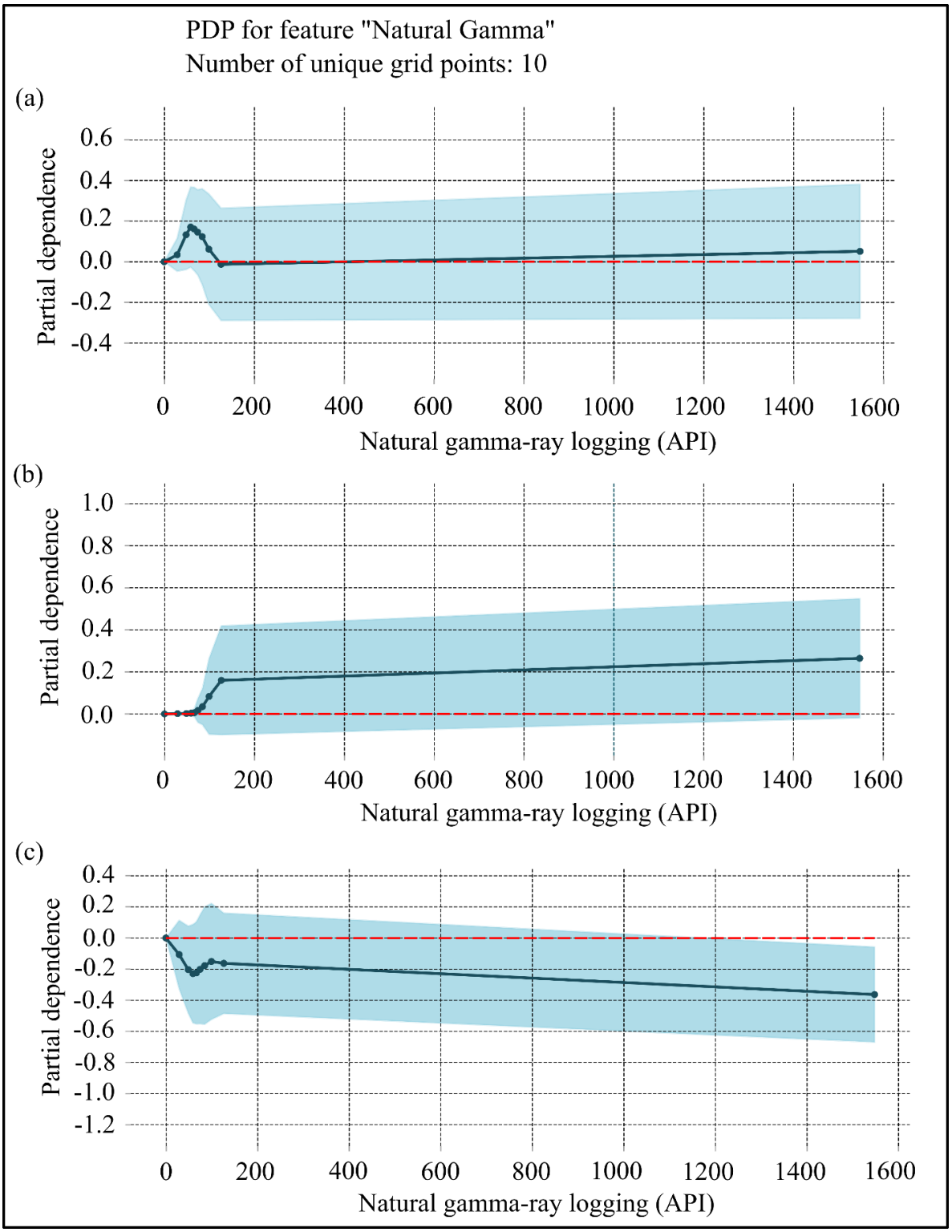

Figure 5.12: Partial dependence plot of gamma-ray log feature in borehole A for Random Forest model classification of: (a) Upper Roodepoort quartzite, (b) Lower Roodepoort quartzite, and (c) Babrosco quartzite.

Also, we evaluated the interactions between the most important logging features (V_p) and other three logging features in the classification of quartzite classes using the Partial Dependence Plot interact (PDP interact). The PDP interact help to elucidate how the logging features interacted to produce the optimal model for the classification of targeted lithology class. **Figure 5.13** shows that the Random Forest model performance in the classification of quartzite classes on the combination of V_p and natural gamma loggings. Note the followings: (1) The model classification performance of Upper Roodepoort quartzite was best when the V_p value $\leq$ 5400 m/s and the gamma ray value was between 0 API and 1400 API is combined, (2) The classification performance of Lower Roodepoort quartzite was best when the value of V_p logging was 5900 m/s, and the value of gamma ray logging was more than 100 API, and (3) The classification of Babrosco quartzite was best when the value of V_p was $\leq$ 5800 m/s and the value of gamma ray logging was 0 API.

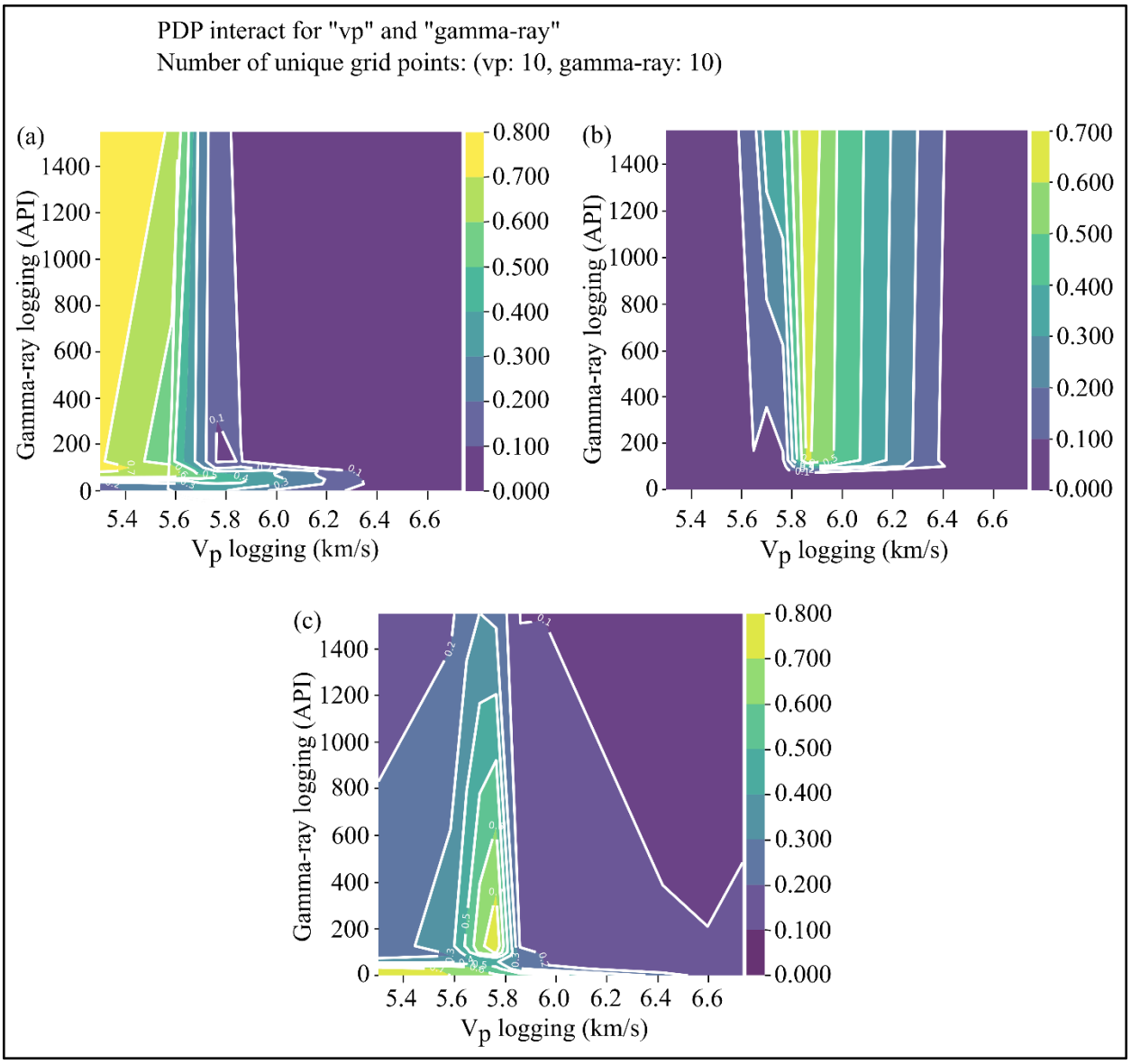

Figure 5.13: Partial Dependence Plot interact between V_p and natural gamma-ray log features in borehole A through Random Forest model for classification of : (a) Upper Roodepoort quartzite, (b) Lower Roodepoort quartzite, and (c) Babrosco quartzite.

Figure 5.14 shows that: (1) Performance of Random Forest model in the classification of Upper Roodepoort quartzite was best when the value of V_p log between 5400mls and 5600 m/s is combined with value of V_s log from 3600 m/s to 3700 m/s, (2) Classification of Lower Roodepoort quartzite was of best performance when the value 5700 m/s of V_p log is combined with the value 3700 m/s of V_s log, and (3) Classification of Babrosco quartzite was of best performance when the

value of V_p log was from 5700 m/s to 5900 m/s and when the value of V_s log was higher than 3800 m/s.

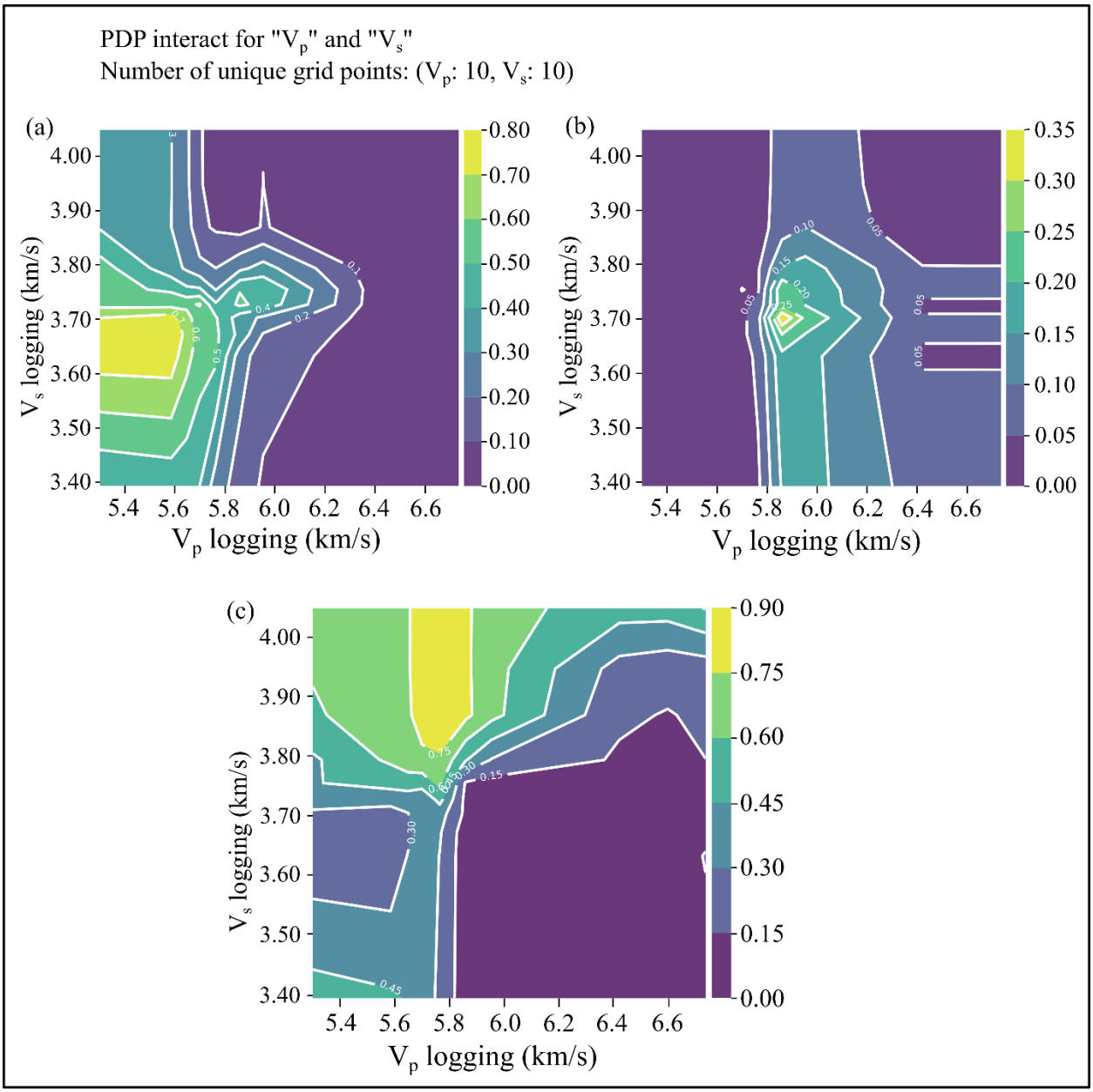

Figure 5.14: Partial Dependence Plot interact between V_p and V_s log features in borehole A through Random Forest model for classification of: (a) Upper Roodepoort quartzite, (b) Lower Roodepoort quartzite, and (c) Babrosco quartzite.

Figure 5.15 illustrates that: (1) When the V_p logging value was less than 5600 m/s and the density logging value was less than 2.8 g/cm^3, Random Forest model performance best in classifying Upper Roodepoort quartzite, (2) When the V_p logging value is 5900 m/s, and the density logging values is less than 2.7 g/cm^3, Lower Roodepoort quartzite classification performance was the best, and (3) Classification of Babrosco quartzite was of best performance when the V_p logging value was between 5700 m/s and 5800 m/s, and the density logging value was between 2.7 g/cm^3 and 3.0 g/cm^3.

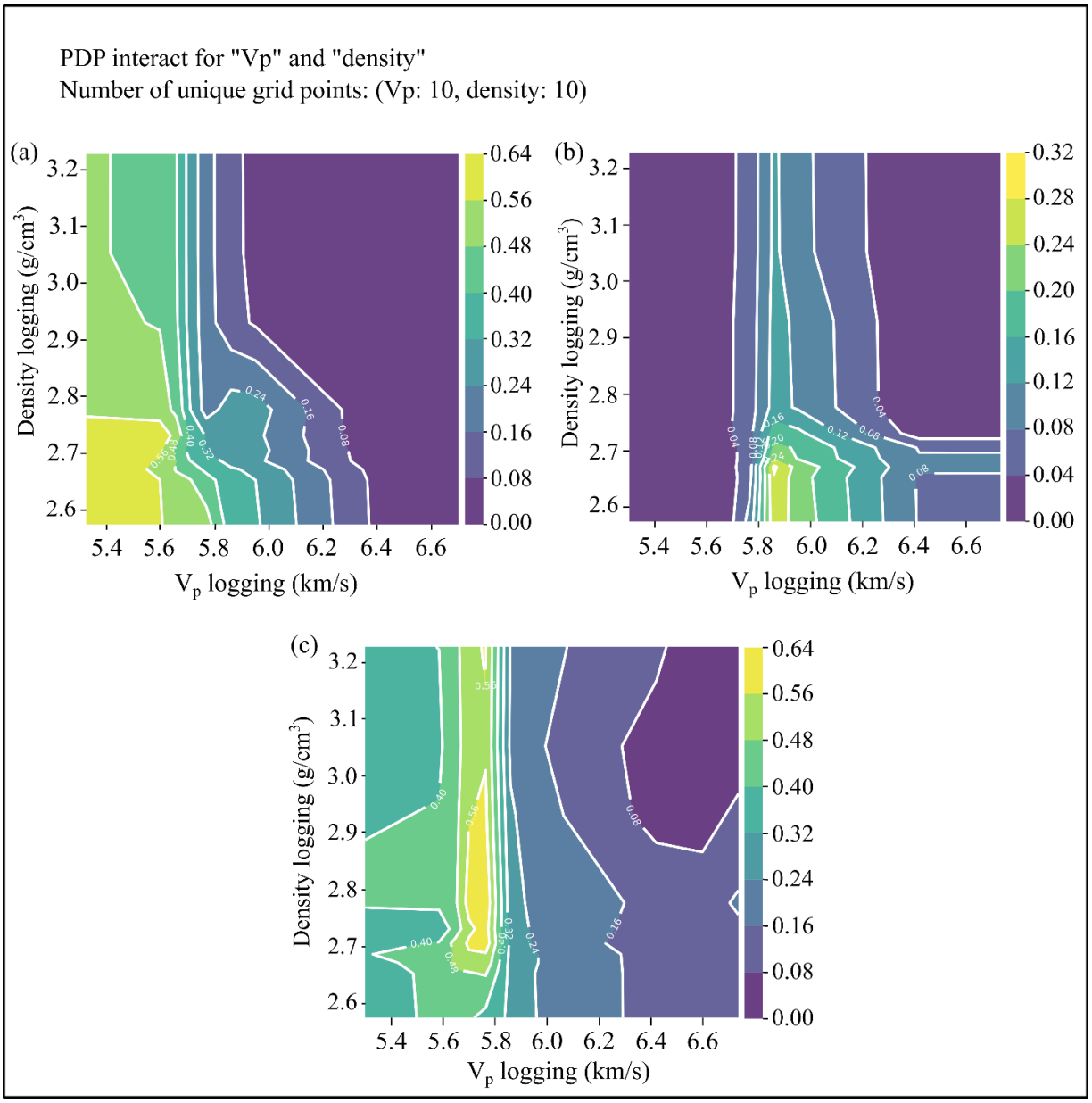

Figure 5.15: Partial Dependence Plot interact between V_p and density log features in borehole A through Random Forest model for classification of: (a) Upper Roodepoort quartzite, (b) Lower Roodepoort quartzite, and (c) Babrosco quartzite.

5.2. Supervised Lithostratigraphic Classification in Borehole B

With our aim to evaluate the correlation of available predicted lithologies between the two boreholes in our study, especially the quartzite classes due to its larger data size, geophysical log datasets from borehole B were analyzed using similar classification models with above discussed workflow in **Figure 4.4**. The dataset was preprocessed applying the same steps discussed in section **4.1.3**. We could not apply datasets from borehole B as the simulated dataset to further evaluate the accuracy and generalizability of borehole A predictive models because it has no similar labelled lithological targets and targets did not fall within the same ranges covered by the training dataset from borehole A. However, borehole B dataset and training dataset from borehole A are of the same origin and logging features.

5.2.1. Optimization and Hyperparameter Tuning

The results of the automatic optimization and hyperparameter tuning processes in borehole B are summarized in **Table 4.9**. **Figure 5.16** and **Figure 5.17** illustrate the results of grid search and 10-fold cross- validation methods application to the Random Forest (RF) classifier model using the hyperparameter listed in **Table 4.1**. **Figure 5.16** showed that for optimal training of RF classifier to avoid overfitting, the applicable parameter search range for number of estimators is [10 , 20, 30, 40, 50, 60, 70, 80, 90,100] trees, maximum depth is [10, 11, 12, 13, 14, 15, 17, 18, 19, 20] and minimum sample split is [2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16,17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29 ,30]. The automatic tuning process of hyperparameters of RF model using the optimal search parameter grids (**Table 4.1** and **Figure 5.16**) and 10-fold cross-validation method yields optimal values of 90, 13, 6, and 0.9084 for number of estimators, maximum depth, minimum samples split, and performance accuracy, respectively (**Figure 5.17**).

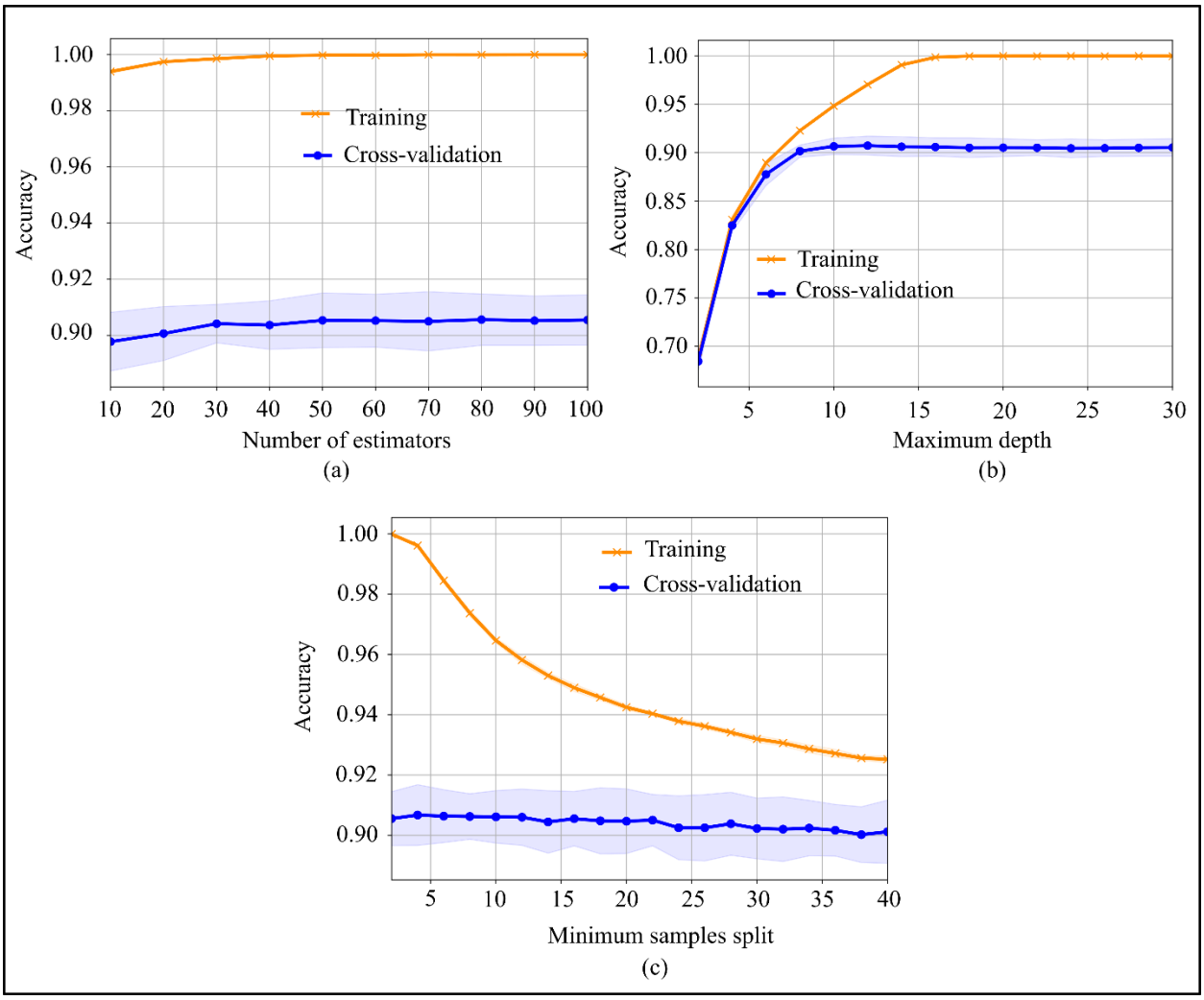

Figure 5.16: Validation curve of gridsearch optimization for Random Forest model's hyperparameters in borehole B: (a) Required trees to construct the forest, (b) Individual Forest maximum growth depth, and (c) Required minimum samples to split an internal node.

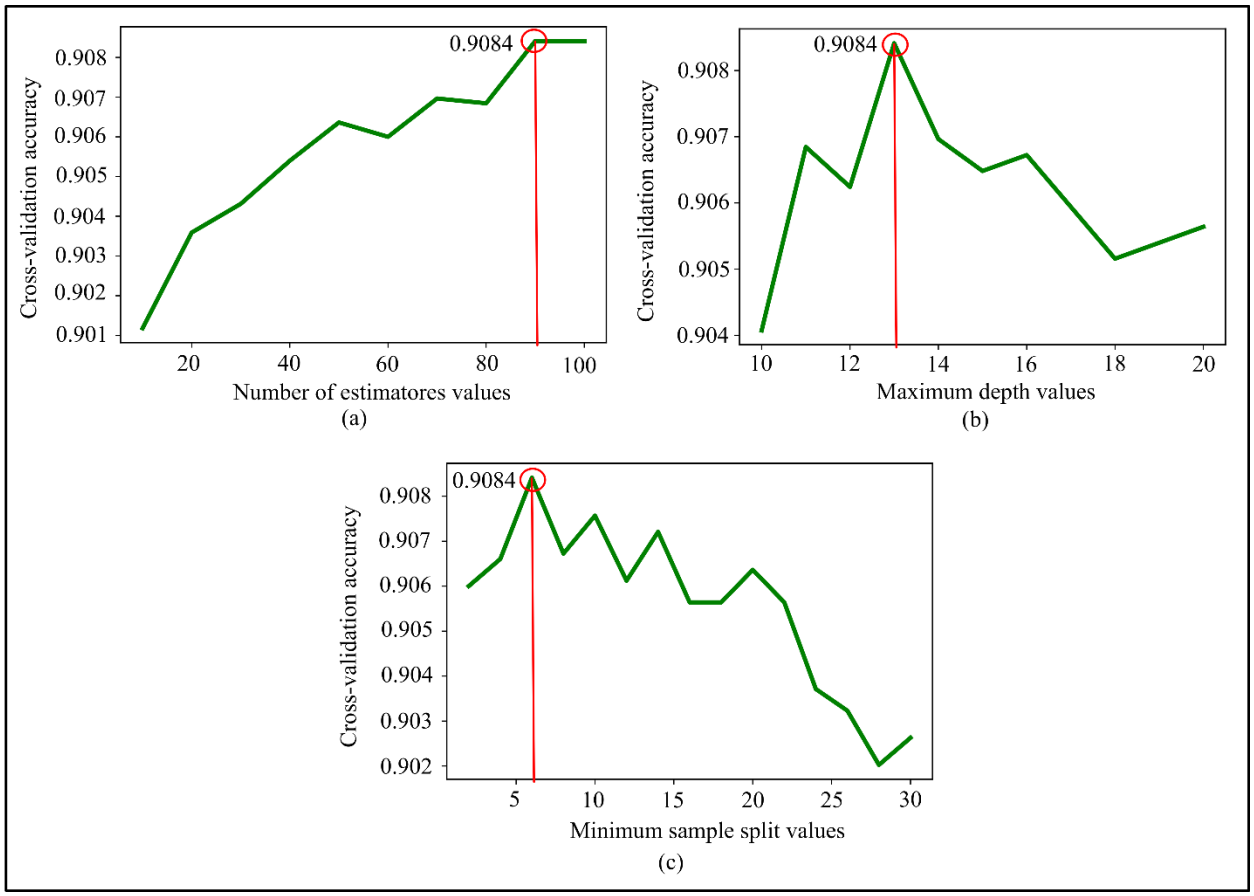

Figure 5.17: 10-fold cross-validation of Random Forest model's hyperparameters in borehole B: (a) Optimal value of number of estimators, (b) Optimal value parameter maximum depth, and (c) Optimal value of minimum samples split.

Applying the appropriate parameter search grid C = [100, 1000, 10000, 100000, 100000], parameter Gamma = [1, 2, 3, 4, 5, 6, 7, 8, 9, 10], and RBF kernel obtained via gridsearch optimization based on the validation curve (**Figure 5.18**) to optimize the Support Vector Machine (SVM) classifier model, we obtained the optimal values of parameter C and Gamma via the 10-fold cross-validation method to be 1000 and 1, respectively. **Figure 5.19** shows that the combination of these optimal parameter values of the SVM model yielded the highest performance accuracy of 0.9147.

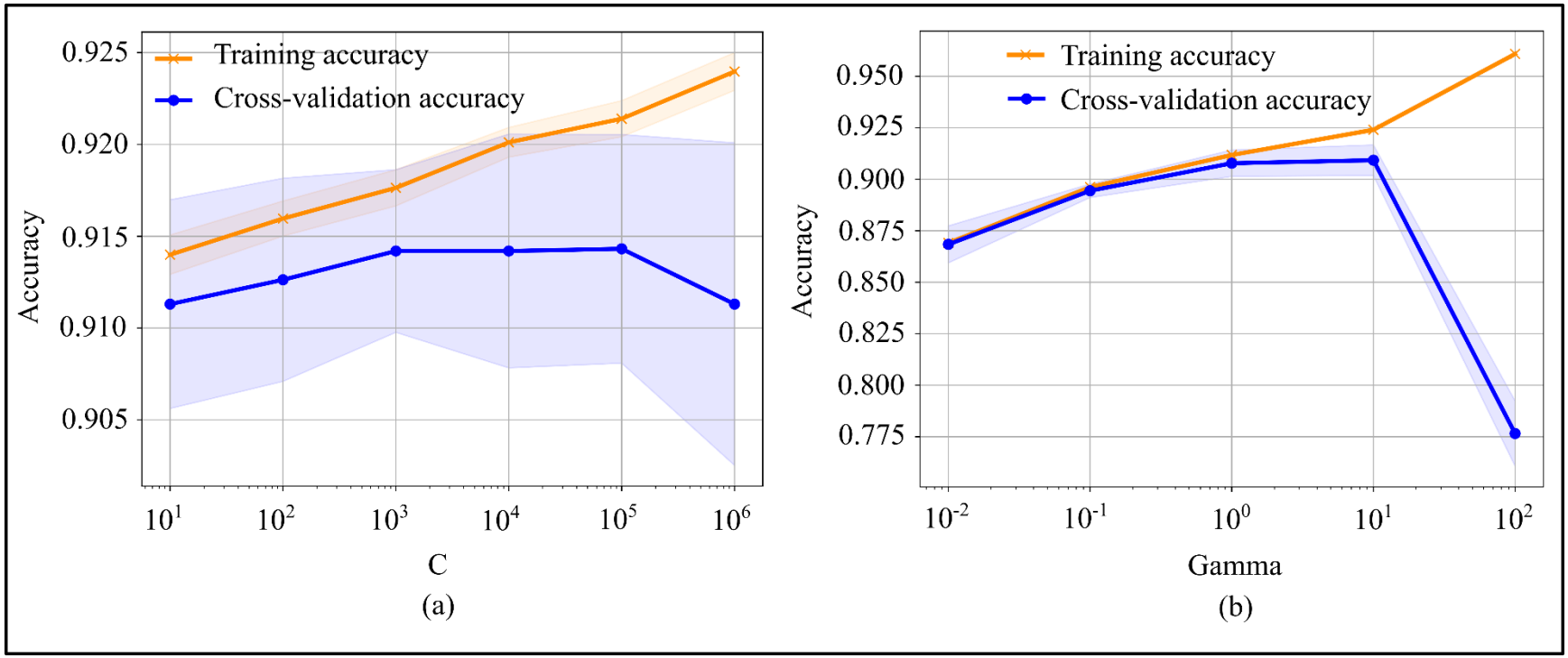

Figure 5.18: Validation curve of gridsearch optimization in borehole B for Support Vector Machine model's hyperparameters: (a) Penalty constant parameter (C), and (b) Kernel spread parameter (Gamma).

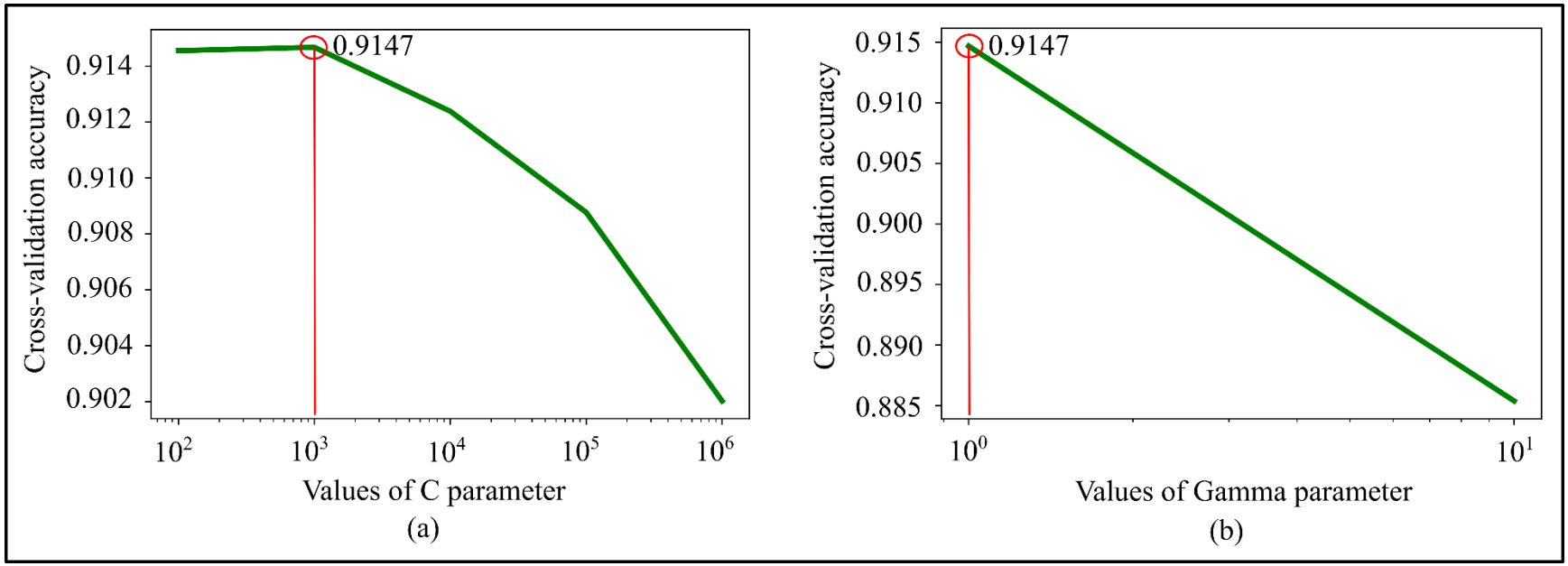

Figure 5.19: 10-fold cross-validation in borehole B for Support Vector Machine model's hyperparameters: (a) Optimal value of the penalty constant parameter (C) when Gamma = 1, and (b) Optimal value of the kernel spread parameter (Gamma) when C = 1000.

For optimal classification performance of Gradient Boosting Decision Trees (GBDT) model, the results of the grid search method shown in **Figure 5.20** demonstrates that the appropriate parameter grid required for number of estimators is [80, 100, 150, 200, 250, 300, 350, 400] trees, learning rate is [0.01, 0.1], minimum samples leaf is [5, 10, 15, 20, 25, 30, 35, 40], minimum samples split is [5, 10, 15, 20, 25, 30, 35, 40], and maximum depth is [5, 10, 15, 20, 25]. **Figure 5.21** illustrates that GBDT model obtained the highest performance accuracy of 0.9065 when the optimal parameter values of number of estimators is 400 trees, maximum depth is 5, minimum samples leaf is 5, learning rate is 0.1, and minimum samples split is 30.

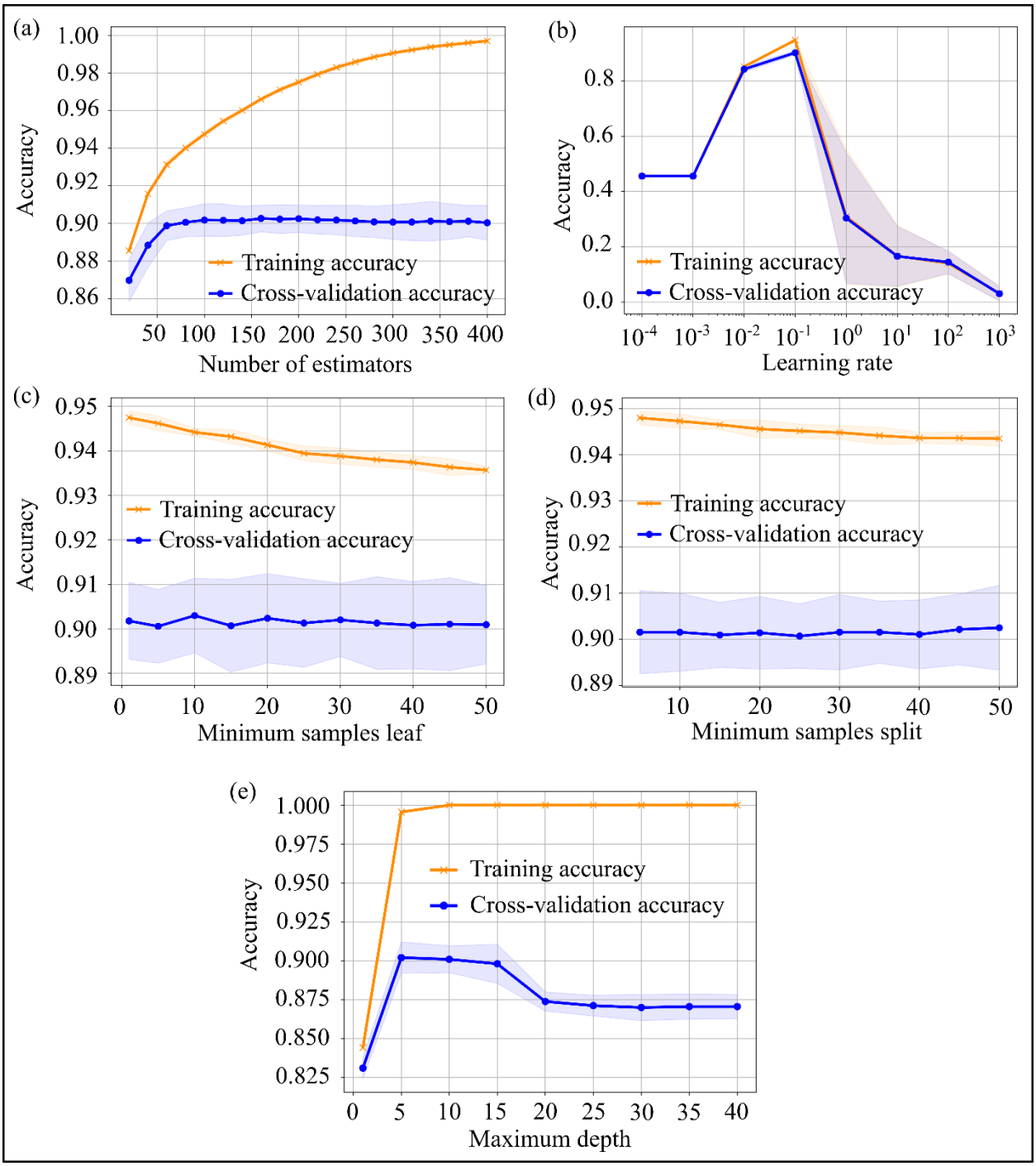

Figure 5.20: Validation curve of gridsearch optimization in borehole B for the GBDT model's hyperparameters: (a) Required number of trees, (b) Learning rate parameter, (c) Required minimum samples at a leaf node, (d) Required minimum samples to split an internal node, and (e) Individual tree maximum growth depth.

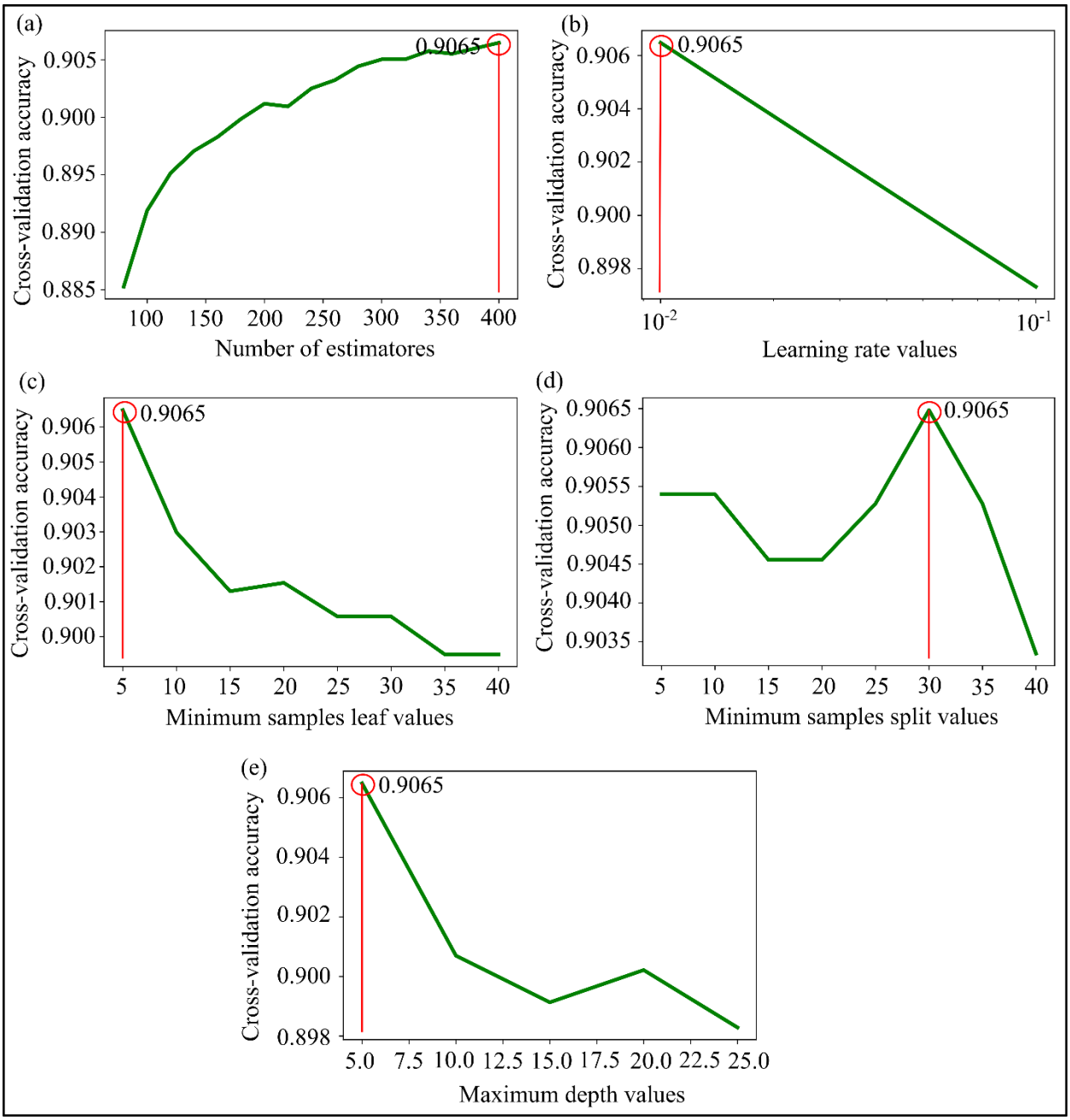

Figure 5.21: 10-fold cross-validation of Gradient Boosting Decision Trees (GBDT) model's hyperparameters in borehole B for: (a) Optimal value of number of trees, (b) Optimal value of learning rate, (c) Optimal value of minimum samples leaf, (d) Optimal value of minimum samples split, and (e) Optimal value of maximum depth.

5.2.2. Performance Evaluation Metrics

To assess the classification performance of evaluated models (RF, SVM and GBDT) to aid cross-comparison for the purpose of selecting optimal model in borehole B, F1 score, precision and recall metrics for every lithological class per evaluated model were estimated based on the training and testing datasets in borehole B. The results of the estimation are listed in **Table 5.4** and **Table 5.5**. Comparing the prediction performance of the evaluated classifiers based on the F1 score on the testing set, the Support Vector Machine (SVM) model is the optimal model with the highest F1 scores of about 0.8953. However, the three evaluated optimized models demonstrated similar classification performance with a difference of about $\leq 1\%$ in their F1 scores (**Table 5.5**). Validating classification per lithological class using the testing dataset, the quartzite classes, Upper Roodepoort siltstone, and Crown basalt were most predicted correctly with F1 score of over 0.90. Upper Roodepoort diabase intrusions were difficult to predict correctly with F1 score $\leq$ 0.27. Support Vector Machine (SVM) model finds it most difficult to identify Upper Roodepoort diabase intrusions (**Table 5.5**).

Table 5.4: Classifier model performance based on each lithological class and several evaluation metrics using **the training datasets** in borehole B. UR and LR denotes Upper Roodepoort and Lower Roodepoort formations, respectively. Weighted average F1 score for a classifier model denoted in red colour.

	Gradient Boosting Decision Tree			Random Forest			Support Vector Machine		
Lithology	Precision	Recall	F1 score	Precision	Recall	F1 score	Precision	Recall	F1 score
UR Quartzite	0.9760	0.9796	0.9778	0.9886	0.9886	0.9886	0.9665	0.9773	0.9719
UR Siltstone	0.9672	0.9964	0.9816	0.9764	0.9976	0.9869	0.9614	0.9940	0.9774
UR Diabase Intrusion	0.8765	0.5917	0.7065	0.9091	0.6667	0.7692	0.8919	0.1375	0.2383
LR Siltstone	0.9017	0.8605	0.8806	0.9473	0.9273	0.9372	0.8838	0.8113	0.8460
LR Diabase Intrusion	0.9032	0.8615	0.8819	0.9516	0.8735	0.9109	0.8671	0.7692	0.8152
LR Quartzite	1.0000	1.0000	1.0000	0.9990	1.0000	0.9995	1.0000	1.0000	1.0000
Crown Basalt	1.0000	1.0000	1.0000	1.0000	0.9873	0.9936	1.0000	1.0000	1.0000
Crown Diabase Intrusion	0.8374	0.9426	0.8869	0.8556	0.9713	0.9098	0.7140	0.9438	0.8130
Weighted average	0.9489	0.9487	**0.9476**	0.9663	0.9654	**0.9648**	0.9276	0.9228	**0.9152**

Table 5.5: Classifier model performance based on each lithological class and several evaluation metrics using **the testing datasets** in borehole B. UR and LR denotes Upper Roodepoort and Lower Roodepoort formations, respectively. Weighted average F1 score for a classifier model denoted in red colour.

	Gradient Boosting Decision Tree			Random Forest			Support Vector Machine		
Lithology Type	Precision	Recall	F1 score	Precision	Recall	F1 score	Precision	Recall	F1 score
UR Quartzite	0.9596	0.9606	0.9601	0.9566	0.9627	0.9597	0.9569	0.9702	0.9635
UR Siltstone	0.9355	0.9457	0.9405	0.9368	0.9674	0.9519	0.9468	0.9674	0.9570
UR Diabase Intrusion	0.3103	0.1500	0.2022	0.4286	0.2000	0.2727	1.0000	0.0667	0.1250
LR Siltstone	0.7740	0.7816	0.7778	0.7929	0.7621	0.7772	0.8201	0.7524	0.7848
LR Diabase Intrusion	0.7899	0.6987	0.7415	0.8043	0.7115	0.7551	0.8421	0.7179	0.7751
LR Quartzite	1.0000	0.9965	0.9982	1.0000	0.9965	0.9982	1.0000	1.0000	1.0000
Crown Basalt	0.9474	0.8571	0.9000	1.0000	0.8810	0.9367	1.0000	1.0000	1.0000
Crown Diabase Intrusion	0.6877	0.8529	0.7615	0.6875	0.8627	0.7652	0.6784	0.9412	0.7885
Weighted average	0.8860	0.8906	**0.8864**	0.8923	0.8954	**0.8913**	0.9145	0.9051	**0.8953**

5.2.3. Confusion Matrix

The confusion matrices for the three implemented classification models (RF, SVM and GBDT) are relatively similar (**Figure 5.22**). To a greater extent, Upper Roodepoort diabase intrusions were difficult to identify. The SVM model could only identify about 0.07 proportion of Upper Roodepoort diabase intrusions, and high proportion of about 0.78 of Upper Roodepoort diabase intrusions were misidentified as Crown diabase intrusions. To a significant extent, the quarzitic classes (over 0.90 fraction), Upper Roodepoort siltstone (over 0.90 fraction), Crown basalt (over 0.80 fraction), and Crown diabase intrusions (over 0.80 fraction) were correctly predicted. Comparison of lithological profile predicted from core logging with machine-learning predicted lithological profile (**Figure 5.23**) reveals that a greater fraction (more than 0.60) of predicted Upper Roodepoort diabase intrusion were classified incorrectly as Crown diabase intrusion, which means that the Upper Roodepoort and Crown formations may have been intruded by intrusions of the same type.

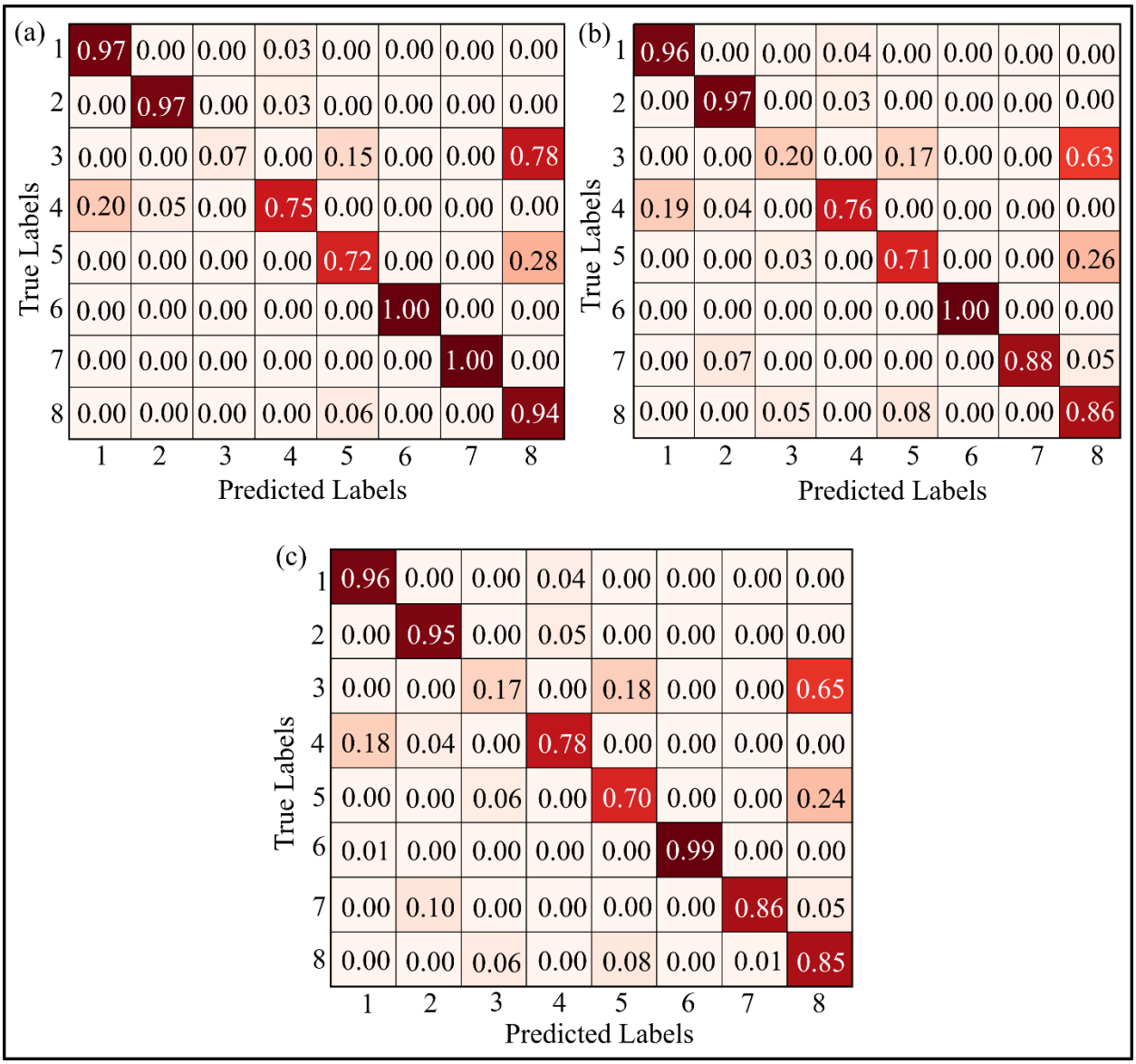

Figure 5.22: Normalized confusion matrices in borehole B based on the **testing datasets** for: (a) Support Vector Machine (SVM) model, (b) Random Forest (RF) model, and (c) Gradient Boosting Decision Trees (GBDT) model. The labels: 0, 1, 2, 3, 4, 5, 6, and 7 denotes Upper Roodepoort quartzite, Upper Roodepoort siltstone, Upper Roodepoort diabase intrusion, Lower Roodepoort siltstone, Lower Roodepoort diabase intrusion, Lower Roodepoort quartzite, Crown basalt, and Crown diabase intrusion, respectively, as encoded in **Table 4.2**.

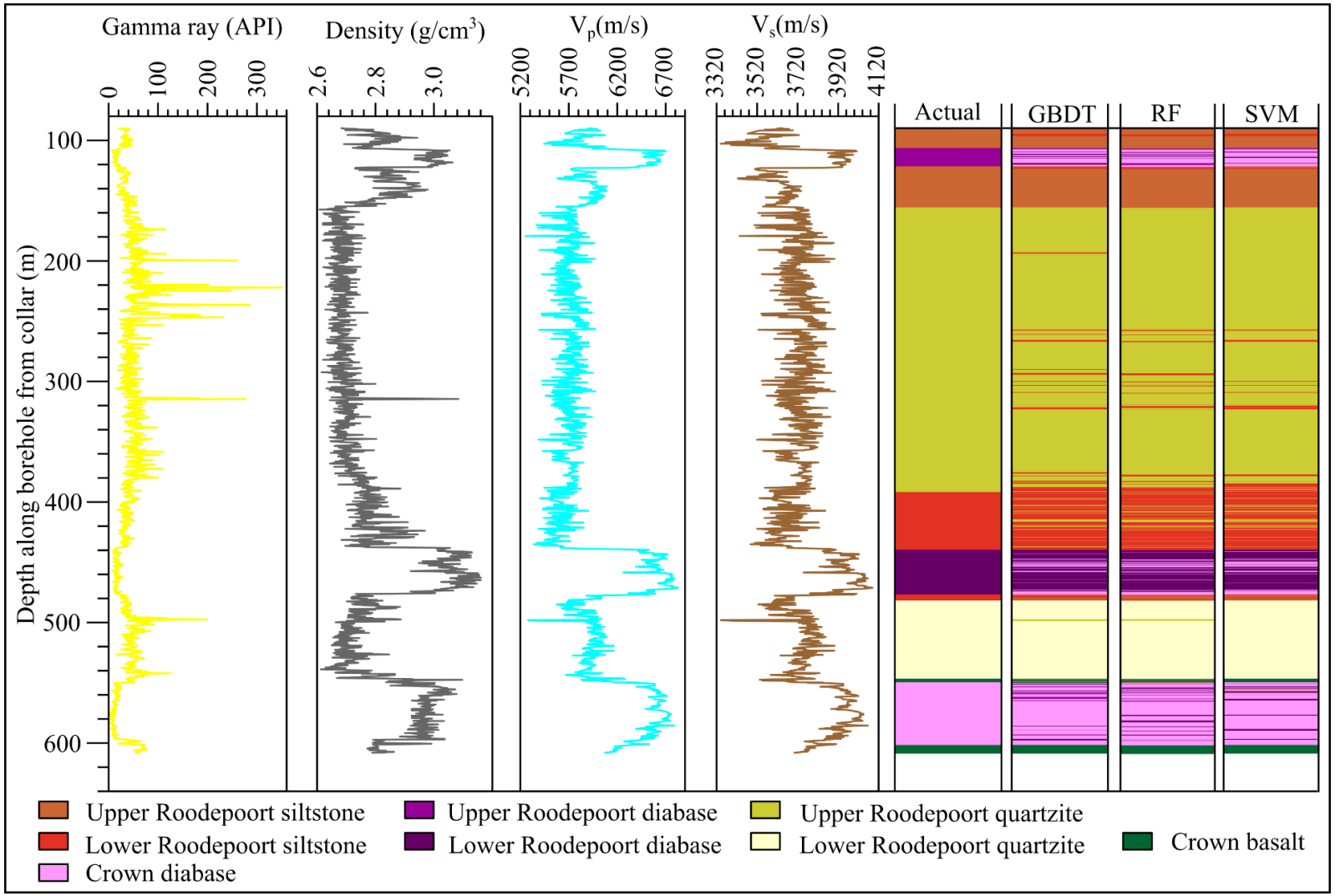

Figure 5.23: Automated lithological profile predicted using geophysical log dataset and machine learning models versus actual lithological profile in borehole B. GBDT is Gradient Boosting Decision Trees, RF is Random Forest, and SVM is Support Vector Machine. Profile sections are colour-coded based on lithologies.

Using Random Forest model, the predictive strength ranking order of features to lithology classification in borehole B using the feature importance process (**Figure 5.24**), (V_p > V_s > density > gamma ray) is different from the ranking order of predictive strength in borehole A (**Figure 5.10**). However, V_p log feature is the most predictive feature to lithology classification in both boreholes.

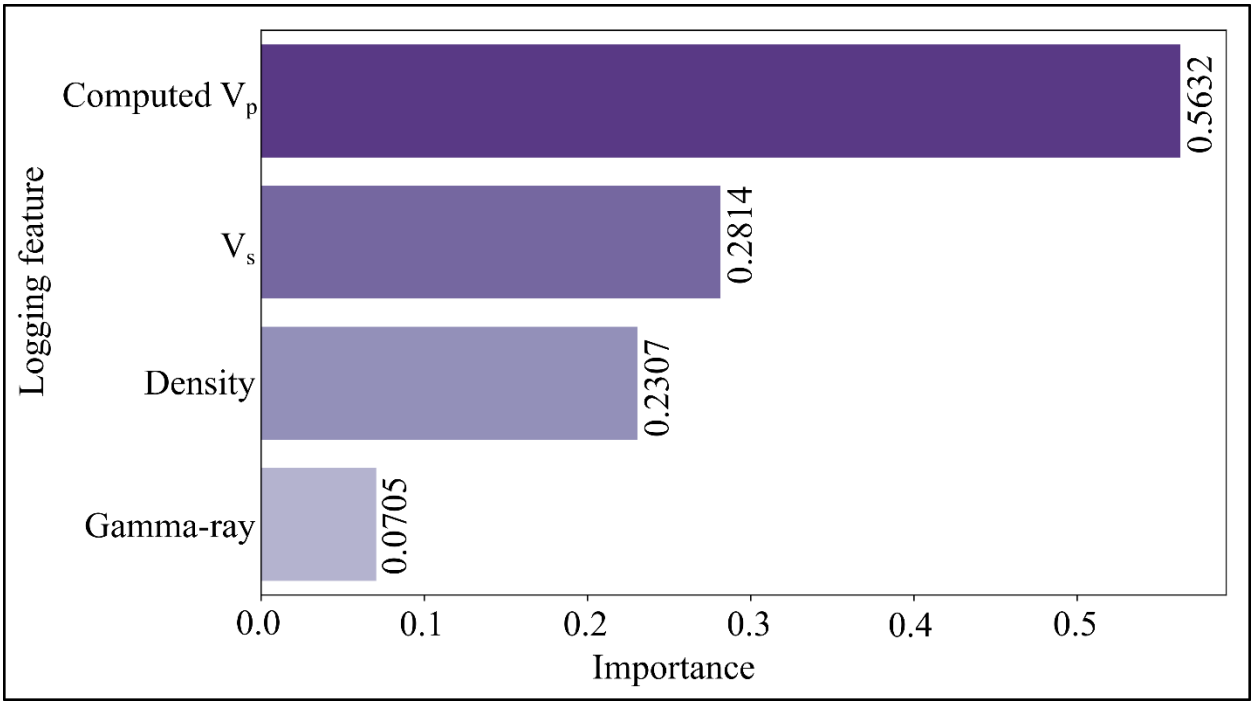

Figure 5.24: Feature importance ranking of log features for Random Forest model (RF) lithology classification in borehole B.

Table 5.6: Logging features importance ranking for Random Forest (RF) model lithology classification in borehole B.

Feature parameter	Feature importance
P-wave log (V_p)	0.5632
S-wave log (V_s)	0.2814
Density log	0.2307
Natural gamma ray log	0..0705

5.3. Impact of Linear Correlation Between Log Features on Classification Performance in Borehole A

5.3.1. Linear Correlation Analysis of Logging Feature Parameters

The linear correlation analysis conducted on the preprocessed logging features in borehole A, presented in **Table 5.7**, indicates that aside from the gamma ray log, the other three collected logging features (V_p, V_s, and density) are correlated with linear correlation coefficients higher than 0.68. V_p and density logs are highly correlated with linear correlation coefficients higher than 0.90. These linear relationships between the recorded log features indicate that linear assumption-based dimension reduction models such as PCA may be applied to transform the logs accurately.

Table 5.7: Correlation analysis between collected logging features.

	Gamma ray log	Density log	P-wave velocity log (V_p)	S-wave velocity log (V_s)
Gamma ray log	1	0.038	0.035	-0.097
Density log	0.038	1	0.921	0.683
P-wave velocity (V_p) log	0.035	0.921	1	0.716
S-wave velocity (V_s) log	-0.097	0.683	0.716	1

5.3.2. Principal Components Analysis (PCA)

Table 5.8 gives the summary of PCA conducted on the preprocessed log features to transform the features into a set of uncorrelated features for the purpose of assessing the impact of correlation between the log features on the classification performance of our models in borehole A. The results (**Table 5.8**) demonstrate that only the first two principal components marked in red colour met the Kaiser (1960) criterion of retaining only Principal Component (PC) having eigenvalues greater than 1.00, while still retaining about 89.29 % of the original information from the dataset. Hence, PC1 and PC2 were used to transform the data. From the values listed in **Table 5.8**, PC1 and PC2 showed eigenvalues of 2.55 and 1.02, respectively. PC1 explains about 63.84% of the original information, and PC2 explains about 25.45% of the original information. Thus, they were retained and fed into the competitive classifier models as input features.

Table 5.8: Summary of PCA results. The red colour denotes PC(s) with eigenvalues greater than 1.00.

Principal Component (PC)	Eigenvalue	Cumulative eigenvalue	Total variance %	Cumulative variance %
1	**2.56**	**2.55**	**63.84**	**63.84**
2	**1.02**	**3.57**	**25.45**	**89.29**
3	0.35	3.92	8.77	99.06
4	0.08	4.00	1.94	100.00

The contribution of logging feature to each principal component was profiled by evaluating the PCA loadings (**Table 5.9**). PCA loadings are impacted by the linear correlations between features and the variability of each feature. **Table 5.9** demonstrates that the V_p is the highest contributor to the PC1 with a coefficient of 95.84%. Meaning that higher V_p results to higher value of the PC1. Also, natural gamma ray contributes negatively to PC1. Meaning that an increase in natural gamma will result in lower values of PC1. The loadings in PC1 agrees with our feature importance results in both Holes A and B, which demonstrates V_p as the most predictive feature to random forest model lithology classification (**Table 5.3**). In PC2, it shows that natural gamma ray is the highest contributor with a coefficient of 99.49% and V_s is the lowest contributor.

Table 5.9: PCA loading of logging features to each principal component.

	PC1	PC2	PC3	PC4
Natural Gamma ray	-0.0087	0.9949	-0.1009	-0.0027
Density	0.9477	0.0703	0.2486	-0.1877
V_p	0.9584	0.0678	0.1921	0.2002
V_s	0.8606	-0.1429	-0.4887	-0.0163

5.3.3. Performance Evaluation Metric

Assessing the impact of the spatial correlation between collected logging features on the classification performance of analyzed classifier models in borehole A, **Table 5.2** and **Table 5.11** suggest that lithologies can be classified using either original or PCA transformed logging data. Similar to the results of lithology classification using the original preprocessed testing dataset (**Table 5.2**), Lower Roodepoort diabase intrusions, Crown basalt, Crown diabase intrusions, and

Babrosco diabase intrusions were most correctly predicted by the PCA transformed testing dataset to an F1 score above 0.70 (**Table 5.11**). To a greater extent, Lower Roodepoort siltstone and quartzite were difficult to identify with F1 score below 0.46. PCA transformed dataset produced Support Vector Machine and Random Forest classifiers as the optimal performing models, which agrees with the prediction results based on application of the original logging dataset. However, the original logging dataset generated more relative accurate prediction results for all the lithology classes and higher performed classifier models with F1 score difference of about 20% (**Figure 5.25**). This suggests that the linear correlation between the logging features has no influence on the lithology classification performance of the models.

Table 5.10: Evaluation metrics per applied classification model on **PCA transformed training dataset** in borehole A. Upper Roodepoort and Lower Roodepoort formations are denoted as UR and LR, respectively. Weighted average of F1score in red colour.

	Gradient Boosting Decision Tree			Random Forest			Support Vector Machine		
Lithology Type	Precision	Recall	F1 score	Precision	Recall	F1 score	Precision	Recall	F1 score
UR Quartzite	0.8942	0.9702	0.9307	0.6881	0.8093	0.7438	0.6598	0.7766	0.7134
LR Siltstone	0.9436	0.8841	0.9129	0.4720	0.2846	0.3551	0.3748	0.2643	0.3100
LR Diabase Intrusion	1.0000	1.0000	1.0000	0.9266	0.8123	0.8657	0.9322	0.8123	0.8682
LR Quartzite	0.9805	0.9685	0.9745	0.6115	0.5035	0.5523	0.5947	0.4283	0.4980
Crown Basalt	0.9805	1.0000	0.9929	0.8070	0.8900	0.8465	0.7815	0.8943	0.8341
Crown Diabase Intrusion	1.0000	1.0000	1.0000	0.8410	0.9452	0.8965	0.8272	0.9404	0.8802
Babrosco Quartzite	0.9493	0.8615	0.9032	0.6521	0.5998	0.6249	0.6102	0.5439	0.5751
Babrosco Diabase Intrusion	1.0000	1.0000	1.0000	0.9876	0.9962	0.9919	0.9839	0.9962	0.9900
Average	0.9470	0.9455	**0.9452**	0.7211	0.7299	**0.7208**	0.6903	0.7001	**0.6909**

Table 5.11: Evaluation metrics per applied classification model on **PCA transformed testing dataset** in Hole A. Upper Roodepoort and Lower Roodepoort formations are denoted as UR and LR, respectively. Weighted average of F1score in red.

	Gradient Boosting Decision Tree			Random Forest			Support Vector Machine		
Lithology Type	Precision	Recall	F1 score	Precision	Recall	F1 score	Precision	Recall	F1 score
UR Quartzite	0.6019	0.7176	0.6547	0.6089	0.7328	0.6651	0.6220	0.7621	0.6850
LR Siltstone	0.3041	0.2407	0.2687	0.3419	0.1852	0.2402	0.3630	0.2454	0.2928
LR Diabase Intrusion	0.7907	0.7514	0.7705	0.8896	0.7569	0.8179	0.8947	0.7514	0.8168
LR Quartzite	0.4697	0.4460	0.4576	0.5254	0.4460	0.4825	0.5429	0.4101	0.4672
Crown Basalt	0.7440	0.7739	0.7586	0.7512	0.8191	0.7837	0.7569	0.8291	0.7914
Crown Diabase Intrusion	0.7459	0.8333	0.7872	0.7565	0.9012	0.8225	0.7629	0.9136	0.8315
Babrosco Quartzite	0.5761	0.4801	0.5238	0.5805	0.5291	0.5536	0.5872	05046	0.5428
Babrosco Diabase Intrusion	0.9926	0.9643	0.9783	0.9928	0.9786	0.9856	0.9964	0.9857	0.9910
Average	0.6384	0.6446	**0.6384**	0.6558	0.6664	**0.6561**	0.6656	0.6741	**0.6645**

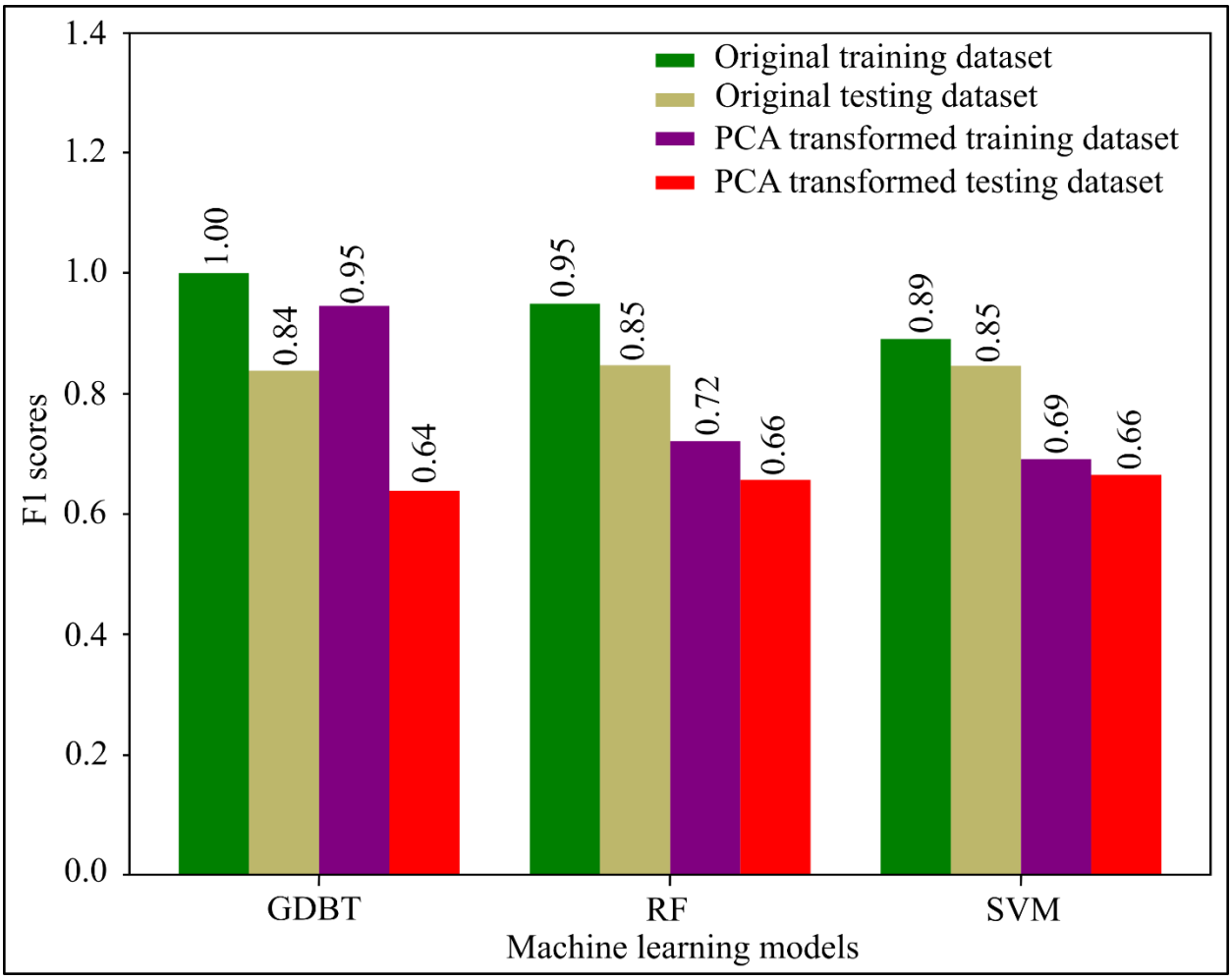

Figure 5.25: Classification performance comparison of the implemented machine learning models in borehole A using original preprocessed dataset with PCA transformed dataset based on the F1 score metric. GBDT is Gradient Boosting Decision Trees, RF is Random Forest, and SVM is Support Vector Machine.

5.4. Correlation Between ICDP DSeis Boreholes

To assess the continuity of quartzite lithologies at relative stratigraphic positions between ICDP DSeis boreholes A and B, the patterns of logs generated by the most important feature in the lithology prediction in both boreholes will be traced. The V_p log, which is the most important feature with less variability in both borehole A and B are traced. Comparison of **Figure 5.9** with **Figure 5.23** proposes that: (1) The quartzite rock formations at the Upper Roodepoort stratigraphic position that was intersected by borehole A at a depth of about 160 to 380 m continued across borehole B at a depth of about 160 to 380 m with V_p logging values ranging from 5300 to 5900 m/s, and (2) The quartzite rock formations at the Lower Roodepoort stratigraphic position that borehole A intersected at a collar depth of 460 to 510 m was crosscut by borehole B at collar depth 490 to 550 m with V_p logging values ranging from 5500 to 6000 m/s. Also, comparing the Partial Dependence Plots (PDP) of the V_p logging for Random Forest model in borehole A with that of borehole B (**Figure 5.11** and **Figure 5.26**), the results demonstrate that: (1) The pattern of the V_p PDP plots for the Upper Roodepoort quartzite rocks in borehole A and B are systematically similar with same curves, and (2) The V_p PDP plots for the Lower Roodepoort quartzite rocks in borehole A and B have similar patterns of curves. In summary, it proved possible to correlate the quartzite strata in borehole A and borehole B because of similarities in their V_p logging feature signatures and because they appear at the same stratigraphic positions.

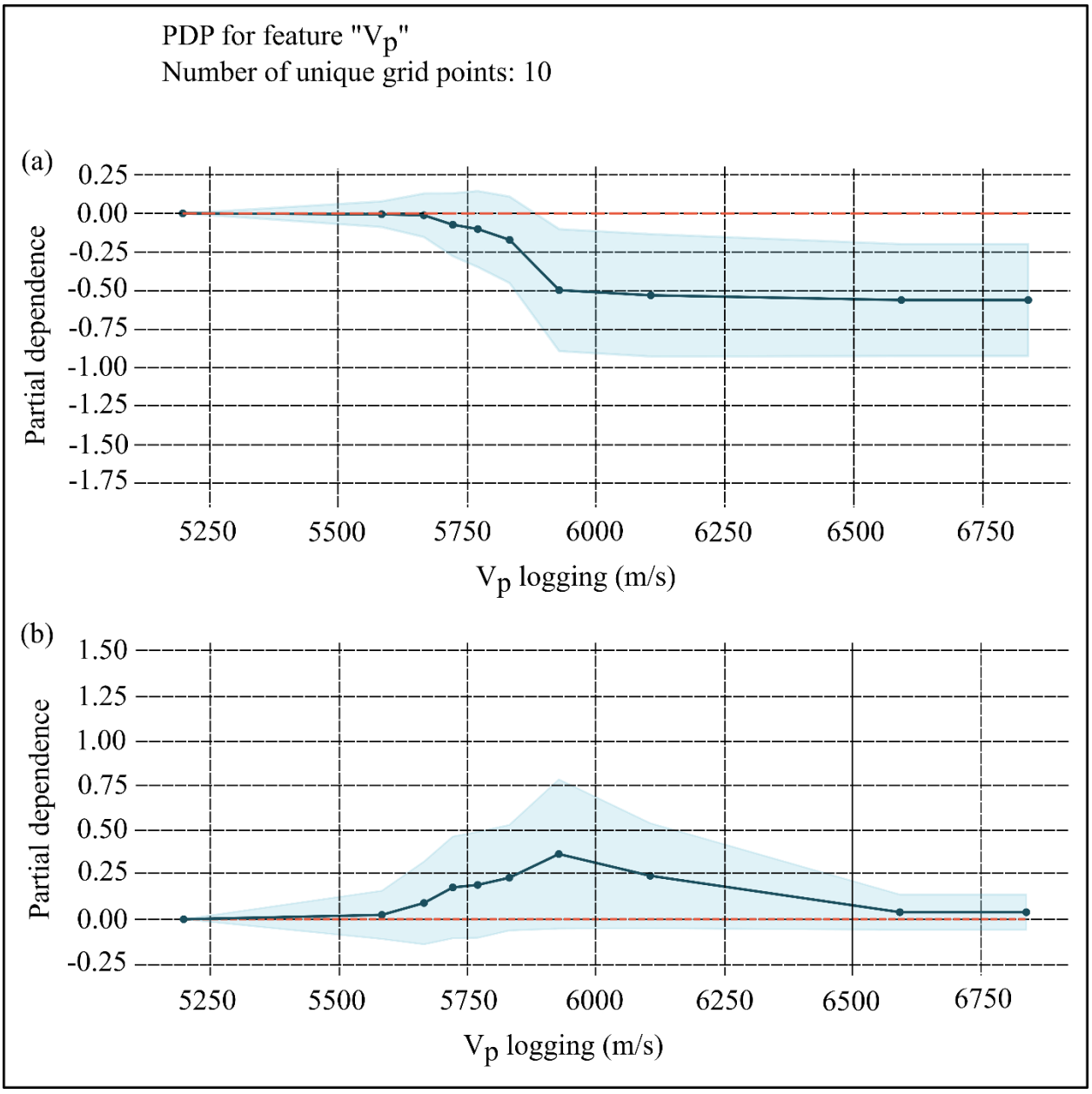

Figure 5.26: Partial dependence plots of V_p log feature for Random Forest (RF) model in borehole B for classification of: (a) Upper Roodepoort quartzite, and (b) Lower Roodepoort quartzite.

5.5. K-means Clustering in Borehole A

5.5.1. Estimation of Optimal Number of Clusters (K)

Computing the optimal number of cluster (K) for the evaluation of K-means clustering using the 'elbow' method and the preprocessed geophysical log datasets from borehole A as discussed in section **4.1.3** and **4.4**, **Figure 5.27** shows that the Sum of Squares starts to decrease below 10000 after the cluster number three. This means that the number of lithology classes expected to be predicted by K-means clustering is three. The optimal choice of three clusters was further verified using the KneeLocator feature from the Skilearn package in Python language (see section **4.1.2**). The predicted classification schemes of three clusters of lithology are different from the traditional geological classifications conducted in the study area , which implies that the logging features of different lithologies demonstrates a significant amount of overlap.

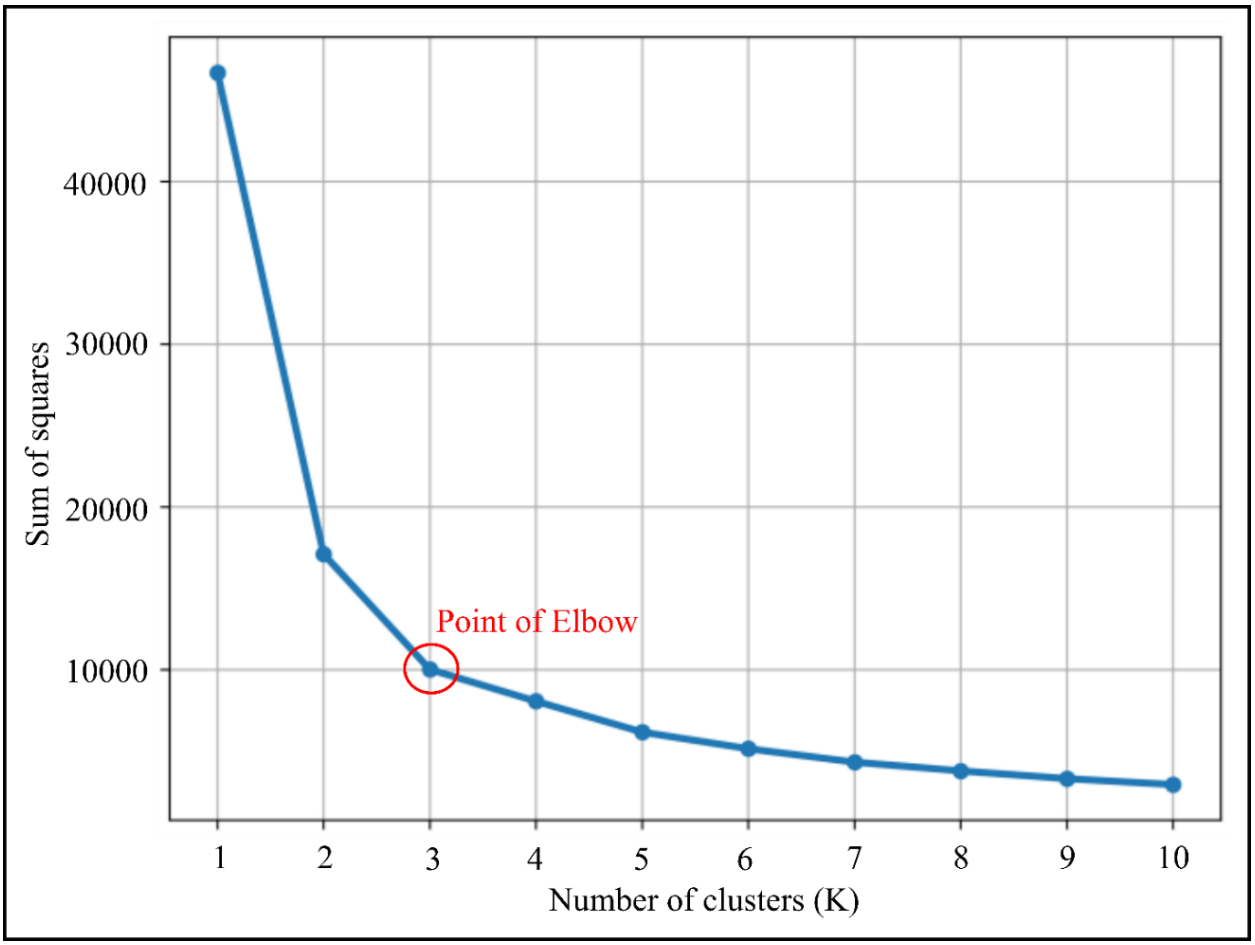

Figure 5.27: Choosing the optimal number of clusters (K) for K-means clustering in borehole A based on the original preprocessed geophysical logs (natural gamma ray, V_p, V_s, and density), and 'elbow' method.

A cross plot of K-means clustering labels shows three clusters of lithology based on their similar response to applied geophysical log features (**Figure 5.28**). A summarized interpretation of the identified clusters representing lithology class is given in **Table 5.12**. The K-means clustering analysis results demonstrate that there are lithological labels (metasiltstone and Crown basalt) that were difficult to identify, which may be due to an overlap in their responses to the evaluated geophysical log features, similar mineral composition, their smaller sample size and heterogeneous nature. For example, the classes of metasiltstone and Crown basalt overlapped with the classes of quartzite. Thus, metasiltstone and Crown basalt were predicted as quartzite. The K-means clustering results were unable to differentiate the different classes of quartzites, which implies that

metasiltstone, Crown basalt, and quartzite is the same lithology type. The produced clusters are colour-coded using lithology label data points predicted by K-means clustering (**Figure 5.28**).

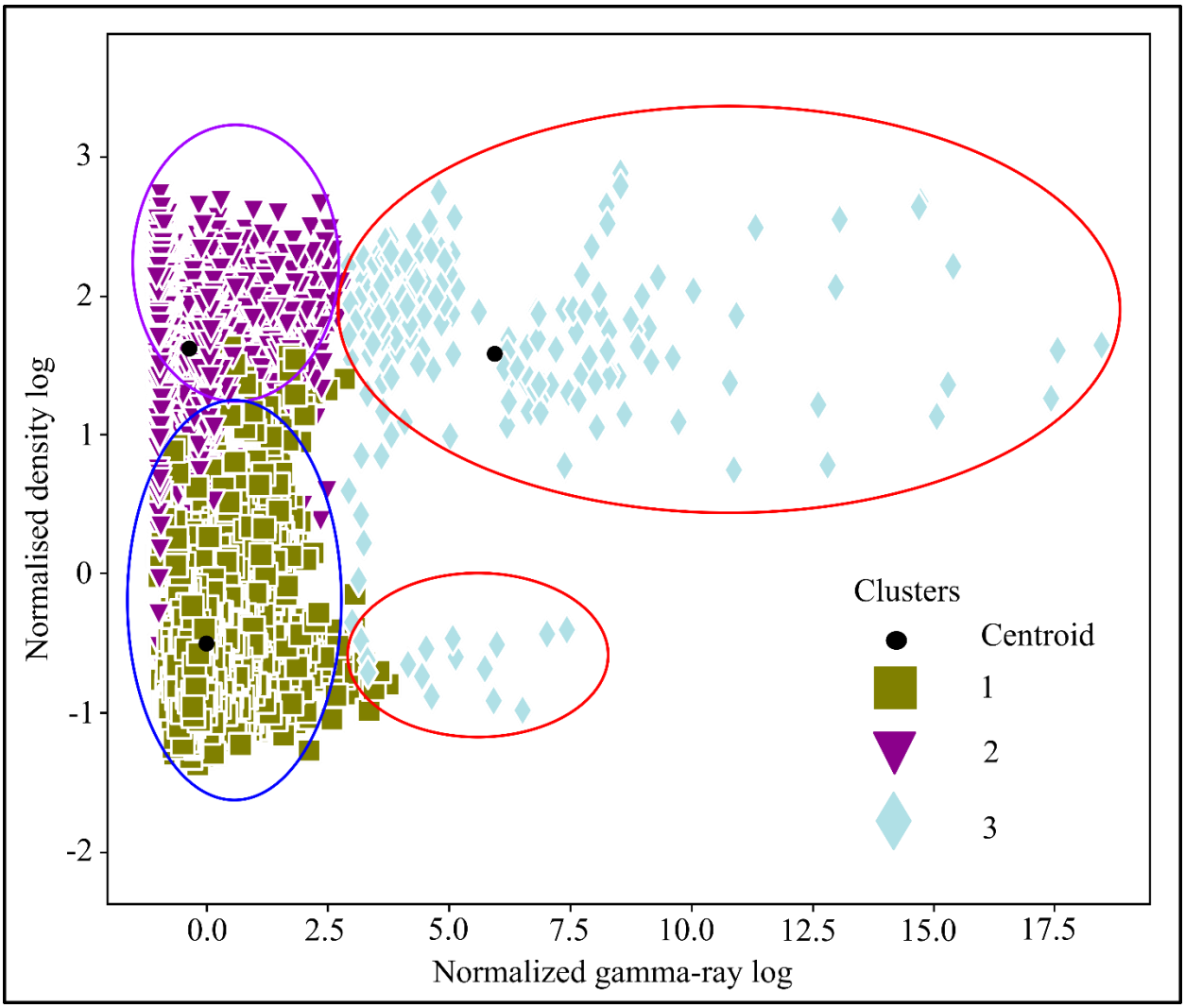

Figure 5.28: Cross plot of density and gamma ray log features in borehole A. clusters are colour-coded based on K-means predicted cluster data points.

Table 5.12: Interpretation of K-means clusters in borehole A. N represents data point counts of clusters.

K-means cluster	Logging feature signatures	Interpretation
N = 9983	Low to high density (2.65 to 2.75 g/cm^3); low to medium gamma ray (0 to 100 API); low to high V_p (5600 to 6200 m/s); and low to medium V_s (3600 to 3850 m/s).	Shaly siltstone and quartzite with clay content and K-feldspar.
N = 2903	Lowest and highest density (2.60 to 3.20 g/cm^3); high to highest gamma ray (600 to 1500 API); lowest and highest V_p (5000 to 6700 m/s); and medium to highest V_s (3850 to 4000 m/s).	Radioactive zones containing potassium elements from high salinity water seepage from fractures, clay size mineral grains such as feldspar and quarts from quartzite and siltstone lithologies.
N = 198	Highest density (3.20 g/cm^3); lowest gamma ray (≤ 0 API); highest V_p (6700 m/s); and highest V_s (4000m/s).	Diabase intrusions (sills and dykes).

Comparing the traditional lithological profile from core logging with K-means clustering predicted lithological profile, a distinct cluster denoted as cluster 3 (as seen in **Figure 5.28**) displaying distinct logging signatures such as: (1) Highest signatures of gamma ray (1500 API), V_p (6700 m/s), V_s (4000 m/s), and density logs (3.20 g/cm^3), and (2) High signature of gamma ray (600 API), lowest signatures of V_p (5500 m/s) and density logs (2.60 g/cm^3), and low V_s (3700 m/s) were identified at two different sections in the Roodepoort strata in borehole A (**Figure 5.29**). These lithostratigraphy locations represents radioactive zones, which may contain potassium elements from high salinity water seeping from a fracture at profile sections 430m, 440, and 465m.

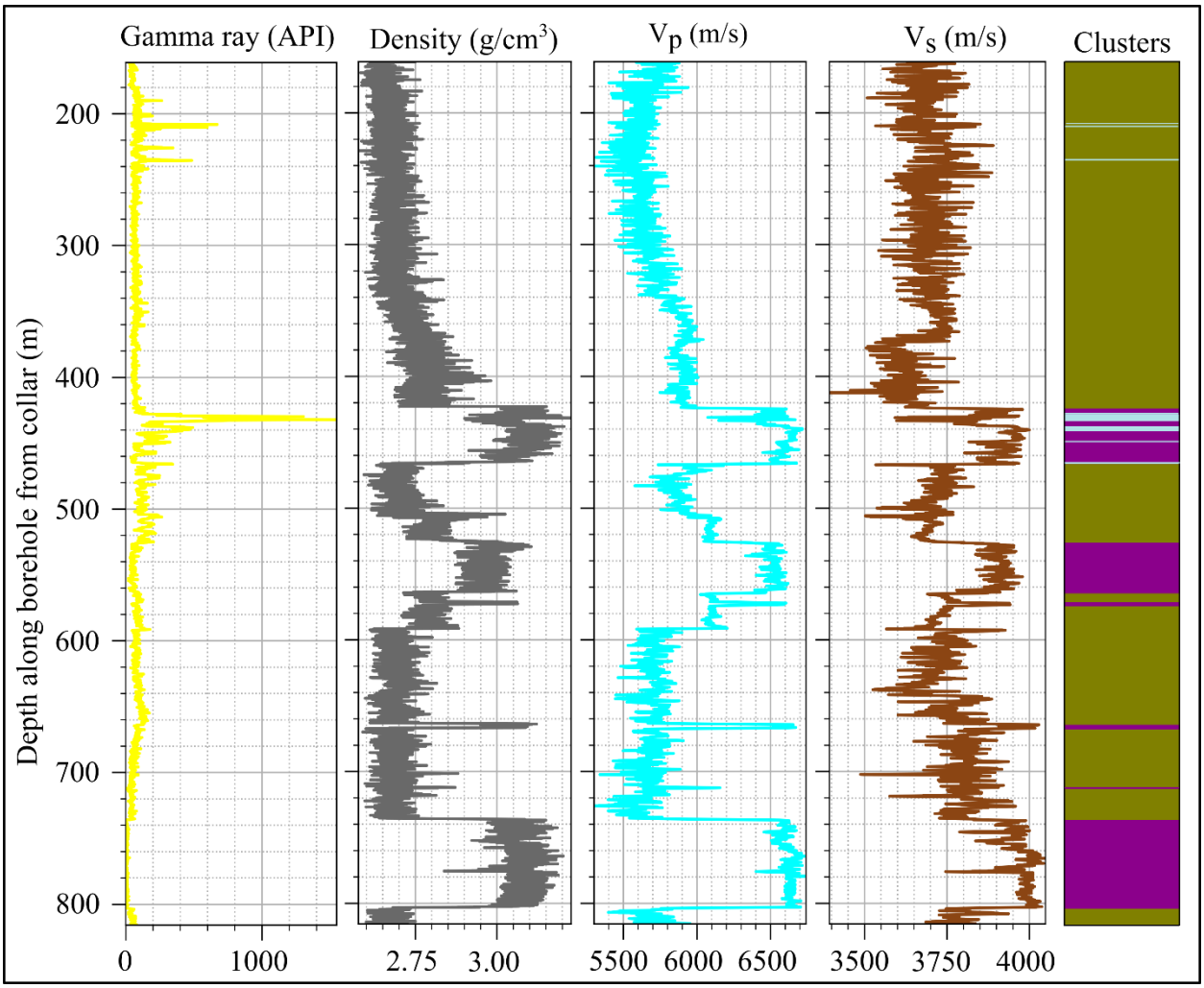

Figure 5.29: Lithological profile predicted by K-means clustering in borehole A using logging dataset. Core sections are based on colour-coded K-means predicted cluster labels.

6. DISCUSSION

6.1. Lithology Classification Through Supervised Machine Learning Models

6.1.1. Optimization and Hyperparameter Tuning Process

The values selected for each hyperparameter of a supervised model can impact classification performance by either improving the accuracy or causing an overfitting, underfitting, or a high misclassification error in the model.

Classification performance of the Random Forest (RF) classifier is highly influenced on how the tree is assembled. The validation score results in **Figure 5.1** and **Figure 5.16** shows: (1) As the level of trees increases from 10, both the training and the cross-validation scores linearly increases. Then, the training and validation scores remain unchanged at a tree level of about 80 and 30 in borehole A and B, respectively, even when the number of trees is increased above 80 and 30 trees to 100 trees in borehole A and B, respectively. This implies that trees ranging between 10 and 100 trees will be optimal for RF accurate prediction. (2) When the maximum growth depth of trees is extremely small (between 1 and 5), underfitting (when the difference between the training score and cross-validation score is small, say a difference < 5%) of the model may occur. But both the training and the cross-validation scores will be unchanged as the depth of trees increases above 18, this means that for accurate prediction of RF model, the optimal value of maximum growth depth of trees should be between 10 and 20. (3) The model may be overfitted (when the difference between the training score and cross-validation score are too large, say > 20%) when the number of samples required for an internal node to split is small, say ≤ 5, so the optimal number of samples required for an internal node to split may be between 6 and 30 for optimal performance of RF model. The application of these appropriate parameter search ranges in boreholes A (**Figure 5.1**) and B (**Figure 5.16**) yielded the optimal values for the RF model parameters and highest performance accuracy of 0.8503 and 0.9084 for borehole A and B, respectively. as illustrated in **Figure 5.17**; **Table 4.8**; and **Table 4.9**.

The results in **Figure 5.3** and **Figure 5.18** shows that larger values of the kernel spread coefficient (Gamma), (i.e., higher than 10^1) and penalty parameter (C), (i.e., higher than 10^6) promotes an increasingly complex hyperplane that may cause overfitting of the Support Vector Machines

(SVM) model. While smaller values of the penalty parameter (C), (i.e., smaller than 10^1) may cause underfitting of the model. Thus, the SVM model will perform better with a smaller (between 10^0 and 10^1) Gamma and a large (between 10^1 and 10^6) C parameters. **Figure 5.4** and **Figure 5.19** demonstrated the results of the combination of **Figure 5.3** and **Figure 5.18**, which presents the optimal values for each hyperparameter of SVM that gave the best classification performance of accuracy 0.8495 and 0.9147 for borehole A and B, respectively. SVM model may have performed slightly better in borehole B due to the smaller data size in borehole B, as SVM model performs better with small dataset.

The learning rate parameter has a substantial impact on the performance of GBDT classifier because it shrinks each decision tree's contribution. For the GBDT model, the results listed in **Figure 5.5** and **Figure 5.20** shows that when the learning rate parameter is too small (between 10^{-4} and 10^{-2}) or too large (between 10^{-1} and 10^3), a minor penalty is introduced on each tree, causing underfitting of the model. This means that for optimized performance of GBDT, the value of learning rate may be 10^{-2} and 10^{-1}. Also, for smaller values of both the number of trees (< 10) and the maximum depth of the individual tree (< 5), the model will underfit. However, neither the training score nor the cross-validation score change at a certain level even when the maximum depth and number of trees are increased above 25 and 120, respectively. This may imply that the optimal values of trees and maximum growth depth required for optimal classification performance of GBDT is between 10 and 400 for number of trees, and 5 and 30 for maximum depth. Changes in the value of minimum samples leaf and minimum samples split have no effect in the training and the cross-validation scores. The optimal values for each hyperparameter of GBDT model that would yield best classification performance accuracy of 0.8461 and 0.9065 for borehole A and B, respectively, through the combination of the results in **Figure 5.5** and **Figure 5.20** is displayed by **Figure 5.6**; **Figure 5.21**; **Table 4.8**; and **Table 4.9**.

6.1.2. Performance Evaluation Metrics

In our study, factors like the model performance metrics (F1 scores) of testing set and computational time of model training proposes only the required optimal model for lithology classification. From the perspective of borehole A, Random Forest (RF) has the highest classification performance metrics (F1 score) of 0.8468, least prediction error, and least training

computational time of about 13.92 hours. Hence, RF may be considered as the best lithology classification model (**Table 4.8**; **Figure 5.7**). This is because in RF classifier model: (1) Trees are built independently using randomly selected features via bootstrap sampling; and (2) Final prediction is the mostly voted prediction output from the testing dataset, which caused the model variance to be low leading it to be more robust to overfitting and have high generalization ability. Also, RF uses fewer iterations while training causing the model to be the fastest.

From the perspective of using the logging features from borehole B as the testing set, the combination of the optimal parameters values suggests the Support Vector Machine model is the optimal model based on the highest F1 score of 0.8953 and least prediction error (**Figure 5.7** and **Figure 5.22**). This is because the Support Vector Machine model is more efficient with small datasets, and, in the case of our study, log data from borehole B is smaller relative to log data from borehole A. However, the Random Forest model demonstrated least computational time (about 5.67 hours) to train () using the available geophysical log data. It is noteworthy that Support Vector Machine (SVM), Random Forest (RF), and Gradient Boosting Decision Trees (GBDT) classification performance metrics varied by just ± 0.0089 F1 scores overall. The significant difference in the computational training time of the models in borehole A and B may be because of a more powerful and faster computer used in implementing the models in borehole B (a desktop with processing Ram 32, Core i9, Windows 10 was used for borehole B analysis, while a desktop with Ram 16, Core i5, and Windows 10 was used for borehole A analysis).

6.1.3. Confusion Matrix

Confusion matrix is used to visualize the prediction performance of machine learning models. Visualizing the lithology prediction performance in borehole A, the confusion matrices shows that the prediction performance of the three evaluated models is relatively similar (**Figure 5.8**). For example, Babrosco diabase intrusion, Crown diabase intrusion, and Crown basalt rocks were much easier to predict by the three models. This may be because they are more homogenous and pure igneous rocks with distinctive highest value of the most predictive log feature (V_p). The significant fraction (about 23.85%) of Babrosco quartzite misclassified as Upper Roodepoort quartzite and vice versa may be due to the great extent of overlap between applied geophysical properties (especially the two most important features that contributed most to the lithology classification:

V_p and gamma ray logs), which may be because of similar mineral composition (quartz grains, plagioclase, pyrite) and similar grain size of the lithologies (**Figure 5.9**). The significant misclassification of Lower Roodepoort quartzite as Upper Roodepoort quartzite (by SVM, RF and GBDT) and Babrosco quartzite (by RF and GBDT) may be due to two main factors (**Figure 5.9**): (1) The geophysical logging signatures (V_p log, gamma ray log, and density log) and grain size of Lower Roodepoort quartzite are highly similar to those of Upper Roodepoort and Babrosco quartzites as a result of similar mineral composition, and (2) Smaller sample size of Lower Roodepoort quartzite, the great variability could make it difficult for the relationship between the geophysical logging signatures and the lithological label to be comprehensively quantified. Hence, the models default to predict dominate lithology with similar log signatures. This agrees with Zhang et al (2021) findings that a significant linear relationship exists between number of samples for a particular label within a dataset and classification accuracy, where the classification accuracy increases with increasing number of samples.

Comparing core section 665.07 – 666.02 m (**Figure 5.9**) of lithological profile predicted using our evaluated machine learning models with the traditional borehole lithological profile from Rickenbacker (2018), Yabe et al (2019), and Nkosi et al (2022) studies, our models' predictions agrees with these studies' results which presented this core section as Upper Roodepoort diabase intrusion. This means that this core section may not be a lithological classification bias by our evaluated classifier models but may actually be a missing labelled data due to human error in documenting the core logging report, which was used in this study as machine learning labelled features.

From borehole B, the confusion matrices results shows that the classification performance of all evaluated machine learning models are relatively similar based on their accuracy scores (**Figure 5.22**). The great misclassification of Upper Roodepoort diabase intrusion as Crown diabase intrusion, and Lower Roodepoort quartzite as Upper Roodepoort quartzite could be due to the following reasons: (1) Great overlap in geophysical logging signatures (mostly density, V_p, and gamma ray loggings) between lithologies due to either similar mineral compositions or grain size, and (2) Upper Roodepoort quartzite and Crown intrusive labels becomes more dominate as their sample sizes are larger compared to the sample size of Lower Roodepoort quartzite and Upper

Roodepoort diabase intrusion, making the relationship between Upper Roodepoort quartzite, Crown diabase intrusion and the signatures of applied geophysical logs to be more strongly defined. Thus, prediction bias arises, and the models' default to mispredict Lower Roodepoort quartzite and Upper Roodepoort intrusive with smaller sample size as Upper Roodepoort quartzite and Crown intrusive, respectively.

In summary, from the models' evaluation results in both borehole A and B, we discover that the lithology classification performance metrics of all the analyzed models are within the same range of F1 scores greater than 0.80. And, regardless of the model used, the most accurate predictions coincide with lithology class that has the largest sample size within the dataset. This demonstrates that the accuracy of machine learning-based classification of lithology depends majorly on the sample size. Nevertheless, Random Forest model has the highest F1 score, least prediction error and least computation training time followed by Support Vector Machine model in borehole A. The Support Vector Machine model has the highest F1 score followed by Random Forest model. This demonstrates the robustness of the Support Vector Machine model to small dataset. While Random Forest takes the least computational time to be trained. Therefore, the choice of optimal machine learning model for lithology classification using geophysical log datasets from our study area may depend on the researcher's specific task case and data size.

With the great spatial variability nature of the applied geophysical datasets in this study, misclassification error and prediction bias are unavoidable despites the automatic optimization and hyperparameter tuning processes conducted. This is because removal of sample classes with fewer data points might not be the best practice especially for geoscientific procedures such as lithology identification and samples mapping, as they might help to detect new unknown geological identifications or discard old and unnecessary identifications done using large data size (Zhang et al., 2021). Also, due to the limited number of boreholes, a large range in data size among the lithology classes, and repeating cycles of lithology types, our models were biased, and lithology misclassification error occurred. However, more accurate work can be done using larger databases, balance data size of lithology classes, more non-overlapping lithology types, and more boreholes to further mitigate misclassification and prediction bias.

The lithological misclassification at the core section (665.07 – 666.02 m) by the three evaluated models (RF, SVM, and GBDT) in borehole A may provide guidance of areas for future lithological studies by geologists, either through core analysis, drilling rock cuttings observations, or laboratory measurements of physical properties to validate the lithology class of this core section.

6.1.4. Predictive Strength of Logging Features to Model Decision Mechanism

Assessing the predictive strength of the applied log features to lithology classification in borehole A using the feature importance process, Random Forest (RF) model was the focus for this assessment because our classification results demonstrates the RF model as the optimal classifier since it has the highest F1 score of about 0.8468 (**Figure 5.7**; **Table 5.2**). As shown in **Figure 5.10** and **Table 5.3**, in terms of feature importance, V_p log ranks first, followed by gamma ray log, V_s log, and finally density logging.

The results in **Figure 5.10** shows the following: (1) The sonic log (compressional wave velocity, V_p) is the most predictive feature to the lithology classification performance of RF model, which can effectively distinguish dense igneous rocks with highest V_p values (such as intrusives rocks in this study) from less dense metasedimentary rocks with frequent interbedding of shale that exhibits lowest V_p values (such as siltstone and quartzites in this study) , (2) Natural gamma ray log is the second most predictive feature. This could help to distinguish lithological classes with highest natural gamma ray values and stable natural gamma ray curves from classes with lowest natural gamma ray values in the study area next. For example, siltstone from the intrusive rocks, (3) Shear wave velocity log (V_s) ranks third in the order of importance. This could aid to classify lithology classes with highest values of V_s and stable curves from classes with lowest V_s values, and (4) Short spaced density logging is the least in importance. This assisted the model in identifying the intrusive rocks that are highly dense and more finely grained from other lithology classes which are less dense and less finely grained.

Additionally, the results of feature importance in **Figure 5.10** and **Table 5.3** could further help to elucidate the misclassification of lithologies across boreholes by the Random Forest model. For example, the values of V_p logging of the different quartzites and siltstones are approximately similar. Hence, the possible misclassification between these classes of lithology. .

Partial Dependence Plots (PDPs) results may assist to elucidate existing correlation between logging features and the targeted quartzite classes classified by the Random Forest model to generate the possibility of accurate classification of targeted lithology. Quartzite classes were our focus of assessment because they are the most dominant with larger sample size. **Figure 5.11** illustrates the partial dependence plots results for the classification of quartzite classes for the V_p log feature: (1) P-wave (V_p) log value of ≤ 5400 m/s made a positive impact on the classification of Upper Roodepoort quartzite. V_p values above 5400 m/s made little impact on the lithology classification, and V_p value above 6400 m/s made no impact on the lithology classification; (2) V_p log value (5400 m/s) made a positive influence on the classification of Upper Roodepoort quartzite. When the value is up to 5600 m/s, the influence of Vp log on the lithology classification increased positively, and when that value is higher than 6400 m/s, V_p log made no changes in the classification of the lithological class; and (3) Vp log value of about 5400 m/s made a positive impact on the classification of Babrosco quartzite. When the Vp value is increased from 5600 m/s to 5800 m/s, the impact on the lithology classification positively increased but starts to decrease when the value is higher than 5800 m/s. When the value reached 6400 m/s, a further increase in V_p log value made no impact. In summary, the PDP shows that: there are not many training data points with large values after 5400 m/s, so the model could not learn a reliable prediction of Upper Roodepoort quartzite for those regions. The model could accurately predict Lower Roodepoort quartzite only from value ranging between 5800 m/s and 6200 m/s, and there are few training data points outside those value ranges. The probability of predicting Babrosco quartzite is unreliable for higher values after 5800 m/s.

Figure 5.12 shows the Partial Dependence Plots (PDPs) for the classification of quartzite classes for the natural gamma ray log feature. The results of **Figure 5.12** shows the following: (1) Value of natural gamma ray log from 0 API to 180 API made an increased positive effect on the classification of Upper Roodepoort quartzite. When the value is from 200 API to 800 API, the predictive strength of natural gamma ray log on the lithology classification decreased. When the value of natural gamma ray reached 1000 API, the contribution to lithology classification gradually increased, while higher natural gamma ray log value above 1400 API made no change in the contribution to Upper Roodepoort quartzite classification, (2) Natural gamma ray from 0 API to 1580 API made positive gradual increase contribution to the classification of Lower Roodepoort

quartzite, and (3) Natural gamma ray value from 0 API to 1580 API made little or no contribution to the classification of Babrosco quartzite. In summary, there are not many training data points at large values beyond 180 API for the model to define the relationship between gamma ray log feature and Upper Roodepoort quartzite for accurate classification. The best possibility of accurately predicting Lower Roodepoort quartzite is high values of gamma ray ranging between 600 and 1600 API. The RF model could not quantify any reliable relationship between gamma ray log and Babrosco quartzite for positive values of gamma ray, as there is no training data point available in those regions.

Figure 5.13 shows the performance of Random Forest model in classifying quartzite classes when the V_p and natural gamma ray log features are combined. Note the followings: (1) The model classification performance of Upper Roodepoort quartzite was optimal when the V_p value was lower than 5400 m/s, and the gamma ray value was higher than 100 API, (2) The classification of Lower Roodepoort quartzite was best when the value of V_p log was 5900 m/s, and the value of natural gamma ray log was more than 100 API, and (3) The classification of Babrosco quartzite was best when the value of V_p was lower than 5800 m/s, and the value of natural gamma ray log was 0 API.

Figure 5.14 shows the performance of Random Forest model in classifying quartzite classes when the V_p and V_s log features are combined. The results of **Figure 5.14** demonstrates the following: (1) Performance of Random Forest model in classifying Upper Roodepoort quartzite was best when the value of V_p log was lower than 5600 m/s, and the value of V_s log was from 3600 m/s to 3700 m/s, (2) Classification of Lower Roodepoort quartzite was of best performance when the value of V_p log was 5700 m/s, and the value of V_s log was 3700 m/s, and (3) Classification of Babrosco quartzite was of best performance when the value of V_p log ranges from 5700 m/s to 5900 m/s, and the value of V_s log was higher than 3800 m/s.

Figure 5.15 illustrates that: (1) When the V_p log values were lower than 5600 m/s, and the density log values were lower than 2.8 g/cm^3, Random Forest model performance in classifying Upper Roodepoort quartzite was the best, (2) When the V_p log value was 5900 m/s, and the density log values were lower than 2.7 g/cm^3, Lower Roodepoort quartzite classification performance was the best, and (3) The Random Forest model was of best performance in the Classification of Babrosco

quartzite when the V_p log values were between 5700 m/s and 5800 m/s, and the density log values ranges between 2.7 g/m^3 and 3.0 g/m^3.

In summary, the higher the importance of a logging feature, the more the predictive strength of the logging feature to lithology classification. The combination of predictive strength of all the logging features to lithology classification contributed to the highest classification accuracy of the Random Forest model. Hence, feature importance results can be utilized to optimize classification modelling and improve lithology classification performance by dropping the least contributing features and focusing on only the most predictive features, which results in reduction of the space dimension of feature causing data density to increase and overfitting to reduce (Nwaila et al., 2022). However, for the purpose of our study, the feature importance evaluation was applied to assess the predictive strength of each logging feature to the classification of lithological labels in order to justify prediction results, which demonstrated that the V_p log is the most vital logging feature required for the accurate classification of lithology in our study area.

The integration of the results of feature importance, PDPs and logging features interaction plots made available an advantageous combination of *in-situ* downhole geophysical logs for the effective identification of lithologies, especially in distinguishing the different classes of the most predominate lithology (quartzite). These could give geologists insights on the most effective geophysical logging feature, and required optimal values that will produce a better lithology classification accuracy (e.g., V_p log at values $\leq$ 5400 m/s for reliable prediction of Upper Roodepoort quartzite by RF model), which in turn saves work time, especially on large data size and study area. The results could serve as control measure for similar future studies.

6.2. Impact of Linear Correlation between Logging Features on Classification Performance

It is important to profile the impact of the linear correlations between logging features on the classification process of machine learning-based models. Performance metrics such as F1 score are used to profile this by comparing the score of each applied machine learning model RF, SVM, and GBDT) based on the preprocessed original log dataset with the PCA transformed log dataset based. Our linear correlation analysis on the applied log features suggests that, aside the natural

gamma ray log, the V_p, V_s, and density logs are strongly correlated with linear correlation coefficients higher than 0.68 (**Table 5.7**). When the classification performance of our applied models based on the preprocessed original log dataset is contrasted with the classification performance based on the PCA transformed dataset (see **Table 5.2**; **Table 5.11**; **Figure 5.25**), it illustrates that the linear correlation between the applied features has no influence on the lithology classification performance of the models. As the F1 scores of models based on the original dataset are relatively higher and the classification of lithologies are more accurate. This may be because machine learning models used in our study are those with nonlinear decision boundaries that appear to treat geometric distortions better than linear decision boundaries models. In summary, our results do not propose that transformation of original data using transformation tools such as Principal Component Analysis (PCA) as preprocessing stage in machine learning workflow prior to the training and evaluation of machine learning-based classification methods is wrong. Instead, transformation of dataset may not be a mandatory stage in the workflow depending on the machine learning algorithm employed, since it does not improve the predictive modelling performance.

6.3. K-means Clustering to Map Radioactive Zones

The prediction of lithological clusters by K-means clustering and the interpretation of the cluster labels based on the responses of the logging features in borehole A is shown in **Figure 5.29** and **Table 5.12**. These results reveal three major lithological cluster labels, which when matched with depth and correlated with logging characteristics did provide useful two lithology classes such as shaly siltstone mixed with quartzite and diabase intrusions. Additionally, when these predicted cluster labels were compared with that of the core logging classes, the K-means clustering method was found useful in extracting extra geological information from the datasets in borehole A. The additional geological information corresponded with the core sections 207.97 – 208.12 m, 210.02 – 210.22 m, 235.32 -235.42, 427.97 – 428.12 m, 429.92 – 433.52 m, 437.42 – 441.22 m, 449.17 - 449.27m (**Figure 5.29**), which could be interpreted as radioactive zones containing radioactive elements based on the high to highest responses of gamma ray log. Finally, core sections 571.12 – 573.67 m, 663.62 – 667.12 m and 711.92 – 712.12 m were extra information of intrusive rocks that might had been missing from the core logging reports due to human error.

Findings from Yabe et al (2019) study showed that core sections 427.97 – 428.12 m, 429.92 – 433.52 m and 437.42 – 441.22 m have seepage of high salinity water from a fracture. This could probably explain the reason for the highest response of all the log features (natural gamma ray, V_p, V_s, and density) at these sections, and may be interpreted to be element of potassium from the salinity water and earth. Other core sections with high natural gamma ray response may be due to the presence of shale and quartzite rocks containing clay size mineral grains such as quartz, feldspar, and chert.

Comparing the lithological profile predicted by K-means clustering as shown in **Figure 5.29** with those predicted by the three (RF, SVM, and GBDT) supervised classification models shown in **Figure 5.9**, the results further validated that the core section 663.62 – 667.12 m may have been intruded by an diabase intrusions, and the misclassification at this core section by the three analyzed supervised models could be a forced misclassification due to error in the training labelled lithology log dataset from the core analysis reports. Especially, when the lithology prediction of this core section (663.62 – 667.12 m) as intrusive rocks by the applied machine learning models also agrees with the study results of Nkosi et al. (2022) and Yabe et al. (2019) lithological profile in borehole A.

In summary, K-means clustering can be applied to effectively classify lithologies into useful clusters based on their similar or dissimilar response to logging feature curves without prior information about the geology. Then, useful classes of lithologies are produced when these clusters are matched with depth and correlated with characteristics of the logging features. These results demonstrated that K-means clustering is able to provide accurate mapping of radioactive zones directly from the combination of natural gamma ray, V_p, V_s, and density logging datasets. Finally, interpreted cluster labels may function as labelled data to train a supervised classifier model and make predictions on new datasets in similar future studies.

It is worth knowing that our K-means clustering results:

(1) Agrees with the results of Random Forest, Support Vector Machine, and Gradient Boosting Decision Tree models that the various quartzite layers are indistinguishable from each other,

(2) Shows that the various intrusives are indistinguishable from each other in borehole A,

(3) Finds it difficult to identify siltstone and Crown basalt layers. Due to large overlap of geophysical properties which may have resulted from composition of similar minerals like shale, plagioclase, pyrite and quartz grains among labels of siltstones, basalts and quartzite, labels of siltstones and Crown basalt were misclassified as labels of shaly quartzite. The overlap of S-wave velocities and bulk densities in the metasedimentary and metabasalts was noted in Nkosi et al. (2022).This may also be due to smaller sample sizes of siltstone and Crown basalt labels compared to larger sample size of quartzite within the applied datasets,

(4) Demonstrates that the basic known lithologies (metasiltstone, quartzite and Crown basalt) in the study area are similar, and

(5) Agrees with the results of Random Forest, Support Vector Machine and Gradient Boosting Decision Tree models that the lithology at borehole core section 663.62 – 667.12 m may be Upper Roodepoort diabase intrusion.

7. CONCLUSIONS

This study aimed to use suite of downhole geophysical log data (V_p, V_s, density and gamma ray) from the ICDP DSeis boreholes at the Moab Khotsong gold mine to assess the applicability of implementing optimized machine learning-based algorithms for classification of lithology. Four machine learning based algorithms were implemented: Random Forest (RF) model, Support Vector Machines (SVM) model, Gradient Boosting Decision Trees (GBDT) model, and K-means clustering. Hyperparameters of RF, SVM and GBDT models were optimized using gridsearch and 10-fold cross-validation methods, while the optimal value of the hyperparameter K of K-means clustering was determined using the "elbow method". The predictive strength of the applied downhole geophysical logs to classification of lithology was assessed using the permutation of covariates through Feature importance analysis and Partial Dependence Plots (PDPs). The classification performance of each model was evaluated using F1 scores and confusion matrices. The original preprocessed dataset was randomly split into training and testing sets in ratio 80:20, respectively, using a Python script in section **4.1.2**.

7.1. Lithology Classification to Differentiate the Different Classes of Quartzites and Intrusives Using Supervised Algorithms

The results show that geophysical log data and machine learning-based algorithms are able to provide accurate predictions of lithology, which could serve as quality guidance for similar future studies both in the industry and academy. When lithological plots predicted by our machine learning-based models are compared with traditional lithological classification plots, it is obvious that our applied machine learning models are accurate. The accurate predictions of the quartzites and diabase intrusion classes illustrate that the accuracy of machine learning-based predictions depends greatly on the sample size of classes within a dataset.

RF model was identified as optimal for lithology classification compared to SVM and GBDT models with highest performance metrics F1 score of 0.8468 and least computational training time for borehole A. The models demonstrated that the different classes of quartzites at different stratigraphical positions are all similar. SVM model was identified as optimal to accurately classify lithologies and show that all the intrusive rocks at different stratigraphic positions are the same

with highest performance metrics F1 score of 0.8953, while RF displayed the least computational training time for borehole B. Therefore, either RF or SVM model may be chosen as an optimal machine learning-based algorithm for lithology classification using geophysical logs in the study area, depending on the researcher's case interest and data size as SVM is more reliable with small data size.

The effective application of feature importance and PDPs plots in profiling the predictive strength of the log features and the interaction between log feature and targeted lithology, respectively, showed that machine learning-based algorithms can also assist geoscientists to optimize time in data analysis by providing the most predictive log feature, its optimal value and optimal interaction between log feature and targeted lithology for improved lithology classification using down hole geophysical logs for objective decision making regarding mine planning and modelling.

The feature importance ranking analysis highlighted that for:

Borehole A (RF model): V_p log > natural gamma-ray log >V_s log > density log, and

Borehole B (RF model): V_p log > V_s log > density log > natural gamma-ray log.

7.2. Impact of Correlation between Logging Features on Classification Performance

We further demonstrated that the linear correlation between our input logging features in borehole A has no significant impact on the lithology prediction performance of our applied models, suggesting that non-learn decision boundaries models handle geometric distortions better than linear decision boundaries models. This implies that the transformation of dataset using transformation tools such as PCA as a preprocessing stage in machine learning-based classification may not be mandatory but depends on the algorithm being used.

7.3. K-means Clustering to Map Radioactive Zones

K-means clustering method effectively mapped radioactive zones possibly containing high content of potassium from high salinity water seeping from a nearby fracture, and clay mineral contents such as feldspar and quart from quartzite and shale interbedded siltstone. K-means clustering reveals that there are three clusters in the study area, which suggests that the basic lithologies (metasiltstone, quartzite and Crown basalt) are similar. These results demonstrated that unsupervised machine learning-based methods can effectively provide extra geological information directly from the datasets (e.g., identification of only two lithologies (shaly quartzite and diabase intrusions) instead of the known four lithologies (metasiltstone, Crown basalt, quartzite, and diabase intrusions) in the study area. This could allow researchers to acquire a new lithology research insight into the research area to validate the similarity between the siltstone, Crown basalt, and quartzite lithologies. Also, the results revealed that K-means clustering can effectively be applied to generate useful clusters of lithological classes, and interpreted cluster labels may function as labels to train a supervised classifier model and make predictions based on geophysical log datasets in similar future studies of lithology classification, especially in cases where the lithology labels are not known or in the case of borehole without full core recovery.

7.4. Limitations

This project was limited by lack of larger database, imbalanced data size among the lithology classes, few boreholes and overlapping lithology types. These contributed to model misclassification and prediction bias. This then suggests that large data and even distribution of data size among classes are key factors to high accuracy of machine learning-based algorithms for classification problems. For future similar work, imbalance data of lithology classes can be addressed using sampling algorithms such as SMOTE, random sampling, and ADASYN (Sevastyanov and Shchetinin, 2020).

REFERENCES

Aha, D. W. and Breslow, L. A. (1998). Comparing simplification procedures for decision trees on an economics classification task. *Naval Research Laboratory Technical Report NRL/FR/5510-98-9881, crosslisted at Navy Center for Applied Research in Artificial Intelligence.*

Ai-min, SH., Zheng, TO. and Jie, GA. (2018). Recognition and Measurement of Pavement Disasters Based on Convolutional Neural Networks. *China Journal of Highway and Transport,* **31**(1), pp.1–10.

Akinyokun, O. C., Enikanselu, P. A., Adeyemo, A. B. and Adesida, A. (2009). Well log interpretation model for the determination of lithology and fluid contents. *The Pacific Journal of Science and Technology*, **10**(1), pp.507-517.

Akkaş, E., Akin, L., Çubukçu, H.E. and Artuner, H. (2015). Application of Decision Tree Algorithm for classification and identification of natural minerals using SEM–EDS. *Computers & Geosciences*, **80,** pp.38-48.

Al-Anazi, A. and Gates, I.D. (2010). On the capability of support vector machines to classify lithology from well logs. *Natural Resources Research*, **19**(2), pp.125-139.

Al-Mejibli, I. S., Alwan, J. K. and Abd Dhafar, H. (2020). The effect of gamma value on support vector machine performance with different kernels. *International Journal of Electrical and Computer Engineering*, **10**(5), pp.5497-5506.

Anscombe, F. J. (1960). Rejection of outliers. *Technometrics,* **2**(2), pp.123-146.

Alves, T. M., Kurtev, K., Moore, G. F. and Strasser, M. (2014). Assessing the internal character, reservoir potential, and seal competence of mass-transport deposits using seismic texture: A geophysical and petrophysical approach. *AAPG Bulletin*, **98**(4), pp.793-824.

Armstrong, R. A., Compston, W., Retief, E. A., Williams, I. S. and Welke, H. J. (1991). Zircon ion microprobe studies bearing on the age and evolution of the Witwatersrand triad. *Precambrian Research*, **53**(3-4), pp.243-266.

Attarchi, S. and Gloaguen, R. (2014). Classifying complex mountainous forests with L-Band SAR and Landsat data integration: A comparison among different machine learning methods in the Hyrcanian Forest. *Remote Sensing*, **6**(5), pp.3624-3647.

Benaouda, D., Wadge, G., Whitmarsh, R. B., Rothwell, R. G. and MacLeod, C. (1999). Inferring the lithology of borehole rocks by applying neutral network classifiers to downhole logs: an example from the Ocean Drilling Program. *Geophysical Journal International*, **136**(2), pp.477-491.

Berset, N. (2017). Characterisation of fractures, lithology and stress state along the first deep borehole drilled at Moab Khotsong gold mine, South Africa. MSc dissertation, Swiss Federal Institute of Technology Zurich, Zurich.

Bougher, B. B. (2016). Machine learning applications to geophysical data analysis. Doctoral dissertation, University of British Columbia.

Breiman, L., Friedman, J. H., Olshen, R.A. and Stone, C. G. (1984). Classification and Regression Trees, Wadsworth International Group, Belmont, California. *Google Scholar*.

Breiman, L. (1996). Bagging predictors. *Machine learning*, **24**(2), pp.123-140.

Breiman, L. (2001). Random forests. *Machine learning*, **45**(1), pp.5-32.

Bressan, T. S. de Souza, M. K., Girelli, T. J. and Junior, F. C. (2020). Evaluation of machine learning methods for lithology classification using geophysical data. *Computers & Geosciences*, **139**, pp.104475.

Catuneanu, O. and Biddulph, M.N. (2001). Sequence stratigraphy of the Vaal Reef facies associations in the Witwatersrand foredeep, South Africa. *Sedimentary Geology*, **141**, pp.113-130.

Chapra, S. C. and Canale, R. O. (1998). Numerical methods for engineers-with programming and software applications. Singapore: WCB.

Chatfield, M. (2017). A bimonthly bulletin for wireline loggers and geoscientists engaged in mining and mineral exploration: a review of the technology, NMR/BMR logging. *Wireline Workshop*, (24)-July.

Chatterjee, R. and Paul, S. (2012). Application of cross-plotting techniques for delineation of coal and non-coal litho-units from well logs. *Geomaterials*, **2**(4), pp.94-104.

Chopra, P., Papp, E. and Gibson, D. (2002). Geophysical well logging. *Geophysical and Remote Sensing Methods for Regolith Exploration*. CRCLEME Open File Report, **144**, pp.105-115.

Cracknell, M.J. and Reading, A.M. (2014). Geological mapping using remote sensing data: A comparison of five machine learning algorithms, their response to variations in the spatial distribution of training data and the use of explicit spatial information. *Computers &* =

Dan, E. and Ijeoma, O. A. (2013). Statistical analysis/methods of detecting outliers in a univariate data in a regression analysis model. *International Journal of Education and Research*, **1**(5), pp.1-24.

Dankert, B. T. and Hein, K. A. A. (2010). Evaluating the structural character and tectonic history of the Witwatersrand Basin. *Precambrian Research*, **177**(1-2), pp.1–22.

Durrheim, R. J. (2015). Geophysical laboratories in deep gold mines: earthquakes, neutrinos, extremophiles and the origin of the Earth's magnetic field. *The History of Geophysics in Southern Africa*. De Beer, J.H. (ed). SUN MeDIA, Stellenbosch, pp.116-138.

Durrheim, R. J. and Riemer, K. L. (2016). The history of mining seismology. *The History of Geophysics in Southern Africa*. De Beer, J. H. (ed). *SUN MeDIA, Stellenbosch*, pp.85-110.

Els, B. G. (2000). Unconformities of the auriferous, Neoarchaean Central Rand Group of South Africa: Application to stratigraphy. *Journal of Africa Earth Sciences*, **30**(1), pp.47-62.

Faweett, T. (2006). Introduction to ROC analysis. *Pattern Recongnition Letters*, **27**(8), pp.861-874.

Fiser-Nagy, A., Varga-Toth, I. and Toth, T. M. (2014). Lithology identification using open-hole well-log data in the metamorphic Kishunhalas-NE hydrocarbon reservoir, South Hungary. *Acta Geodaetica et Geophysica*, **49**(1), pp.57-78.

Forgy, E. W. (1965). Cluster analysis of multivariate data: efficiency versus interpretability of classifications. *Biometrics*, **21**(3), pp.768-769.

Friedman, J. H. (2002). Stochastic gradient boosting. *Computational statistics & data analysis*, **38**(4), pp.367-378.

Frimmel, H. E. and Gartz, V. H. (1997). Witwatersrand gold particle chemistry matches model of metamorphosed, hydrothermally altered placer deposits. *Mineralium Deposita*, **32**(6), pp.523-530.

Frimmel, H. E. and Minter, W. E. L. (2002). Recent developments concerning the geological history and genesis of the Witwatersrand gold deposits, South Africa. *Society of Economic Geologists, Special Publications*, **9**, pp.45-117.

Frimmel, H. E., Groves, D. I., Kirk, J., Ruiz, J., Chesley, J. and Minter, W. E. L. (2005). The formation and preservation of the Witwatersrand goldfields, the world's largest gold province. *Society of Economic Geologists 100th Anniversary Volume,* **769**, pp.767.

Fushiki, T. (2011). Estimation of prediction error by using K-fold cross-validation. *Statistics and Computing*, **21**(2), pp.137-146.

Garcia, M.H., Rabaute, A., Yven, B. and Guillemot, D. (2011). Multivariate and spatial statistical analysis of Callovo-Oxfordian physical properties from lab and borehole logs data: Towards a characterization of lateral and vertical spatial trends in the Meuse/Haute-Marne Transposition Zone. *Physics and Chemistry of the Earth, Parts A/B/C*, **36**(17-18), pp.1469-1485.

Gelfort, R. (2006). On classification of logging data. Doctoral dissertation, master's thesis, Claustra University of Technology, Germany.

Glover, P.W. (2000). Petrophysics. *University of Aberdeen, UK*, **270**.

Graham, J. W. (2009). Missing data analysis: Making it work in the real world. *Annual Review of Psychology*, **60**(1), pp. 549-576.

Grinand, C., Rakotomalala, F., Gond, V., Vaudry, R., Bernoux, M. and Vieilledent, G. (2013). Estimating deforestation in tropical humid and dry forests in Madagascar from 2000 to 2010 using multi-date Landsat satellite images and the random forests classifier. *Remote Sensing of Environment*, **139**, pp.68-80.

Gueguen, Y. and Palciauskas, V. (1994). Introduction to the physics of rocks. Princeton University Press.

Guiwen, W., Zhongchun, S., Jianwei, F., Xingping, L., Xianling, Z. and Tuo, P. (2015). Control factors and logging evaluation method for glutenite reservoir in Mabei Area, Junggar Basin. *Xinjiang Petroleum Geology*, **36**(1), p.1.

Gullo, F. (2015). From patterns in data to knowledge discovery: What data mining can do. *Physics Procedia*, **62**, pp.18-22.

Guoming, L. (2010). Logging comprehensive identification technology of deep sandy conglomerate lithology, dongying sag. *Well logging technology*, **34**(2), pp.168-171.

Hamel, L. (2009). Knowledge Discovery With Support Vector Machines, A John Wiley & Sons. *Inc., Hoboken, New Jersey*, pp.246.

Hamer, R. M. and Simpson, P. M. (2009). Last observation carried forward versus mixed models in the analysis of psychiatric clinic trials. *American Journal of Psychiatry*, **166**(6), pp. 639-641.

Han, Q., Zhang, X. and Shen, W. (2018). Lithology identification technology based on gradient boosting decision tree (GBDT) algorithm. *Bulletin of Mineralogy Petrology and Geochemistry*, **37***(06)*, pp.1173-1180.

Hastie, T., Tibshirani, R., Friedman, J.H. and Friedman, J.H. (2009). The elements of statistical learning: data mining, inference, and prediction. *New York: springer*, **2**, pp. 1-758.

Hayat, U., Ali, A., Murtaza, G., Ullah, M., Ullah, I., de Celis, A. N. and Rajpoot, N. (2019). Classification of well log data using vanishing component analysis. *Pure and Applied Geophysics*, **177**(6), pp. 2719-2737.

Hawkins, D. M. (1980). Identification of outliers. 11, London: Chapman and Hall.

Horsfall, O. I. and Davies, O. (2016). Characterization of reservoir fluid using body wave velocities. *Nigerian Journal of Oil and Gas Technology*, **2**(1), pp. 21-31.

Hossin, M. and Sulaiman, M.N. (2015). A review on evaluation metrics for data classification evaluations. *International Journal of Data Mining & Knowledge Management Process*, **5**(2), p.1.

Hsieh, B., Lewis, C. and Lin, Z. (2005). Lithology identification of aquifers from geophysical well logs and fuzzy logic analysis: Shui-Lin Area, Taiwan. *Computer & Geosciences*, **31**(3), pp.263-275.

Hsu, Chih-Wei and Lin, Chih-Jen. (2002). A comparison of methods for multiclass support vector machines. *in IEEE Transactions on Neural Networks*, **13**(2), pp.415-425.

Huang, B. and Pan, B. (2004). Characteristics of log responses and lithology determination of igneous rock reservoirs. *Journal of Geophysics and Engineering*, **1**(1), pp.51-55.

Ishida, A., Ogasawara, H. and Sugimura, K. (2018). Stress-state inferred from DCDA measurements at Moab Khotsong gold mine., Ritsumeikan University, Kyoto.

Jahdhami, N. A. and Anboori, A. A. (2017). The application of specific drilling energy to identify overburden lithological boundaries and aid well operations-Oman Khazzan field. *In Abu Dhabi International Petroleum Exhibition & Conference, Society of Petroleum Engineers*.

Jia, D. and Xue, R. (2022). Research on earnings management of growth enterprise market in China Stock Market: comparative analysis based on the BPNN, GBDT, and MLR models. *Computational Intelligence and Neuroscience*.

Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Educational and Psychological Measurements*, **20**(1), pp.141-151.

Kang, H. (2013). The prevention and handling of the missing data. *Korean Journal of Anesthesiology*, **64**(5), pp.402-406

Kazmierczuk, M. and Jarzyna, J. (2006). Improvement of lithology and saturation determined from well logging using statistical methods. *Acta Geophysica*, **54**(4), pp.378-398.

Kearey, P., Brooks, M. and Hill, I. (2002). An introduction to geophysical exploration, 3rd edition. Great Britain: Blackwell Science Ltd, 272.

Kecman, V. (2005). Support vector machines–an introduction. In *Support vector machines: theory and applications*, pp.1-47. Springer, Berlin, Heidelberg.

Killeen, P.G. (1997). Borehole geophysics: exploring the third dimension. In *Proceedings of Exploration*, **97**, pp. 31-42.

Kim, J. O. and Curry, J. (1977). The treatment of missing data in multivariate analysis. *Sociological Methods & Research*, **6**(2), pp. 215-215.

Konaté, A.A., Pan, H., Ma, H., Cao, X., Ziggah, Y.Y., Oloo, M. and Khan, N. (2015). Application of dimensionality reduction technique to improve geophysical log data classification performance in crystalline rocks. *Journal of Petroleum Science and Engineering*, **133**, pp.633-645.

Krygowski, D. A. (2003). Guide to petrophysical interpretation. Austin, Texas, USA.

Kuhn, S., Cracknell, M. J. and Reading, A.M. (2018). Lithologic mapping using random forests applied to geophysical and remote-sensing data: A demonstration study from the Eastern Goldfields of Australia. *Geophysics*, **83**(4), pp. B183-B193.

Lai, J., Wang, G. and Zheng, Y. (2013). Types and interpretation methods of lithology and lithofacies of the second member of Xujiahe Formation in Penglai Area, Central Sichuan Basin. *Duankuai Youqitian (Fault-Block Oil & Gas Field)*, **20**(1), pp.33-37.

Law, J. D. M., Balley, A. C., Cadle, A. B., Phillips, G. N. and stanistreet, I. G. (1990). Reconstructive approach to the classification of Witwatersrand 'quartzites'. *South African Journal of Geology*, **93(**1), pp. 83–92.

Li, Y. and Anderson-Sprecher, R. (2006). Facies identification from well logs: A comparison of discriminant analysis and naïve Bayes classifier. *Journal of Petroleum Science and Engineering*, **53**(3-4), pp.149-157.

Li, C., Wang, J., Wang, L., Hu, L. and Gong, P. (2014). Comparison of classification algorithms and training sample sizes in urban land classification with landsat thematic mapper imagery. *Remote sensing*, **6**(2), pp.964-983.

Lloyd, S. P. (1982). Least squares quantization in PCM. *Institute of Electrical and Electronics Engineers Transaction on Information's Theory*. 2**8**(2), pp. 129-137.

Lofts, J.C. (1993). Integrated geochemical-geophysical studies of sedimentary reservoir rocks. Thesis PhD, University of Leicester (United Kingdom).

Luthi, S. (2001). Geological well logs: Their use in reservoir modeling. Springer Science & Business Media.

Mahesh, B. (2020). Machine learning algorithms-a review. *International Journal of Science and Research (IJSR). [Internet]*, **9**, pp.381-386.

Mahmoodi, O. and Smith, R. (2015). Clustering of downhole physical property measurements at the Victoria property, Sudbury for the purpose of extracting lithological information. *Journal of Applied Geophysics*, **118**, pp.145-154.

Mahmoodi, O., Smith, R. S. and Tinkham, D. K. (2016). Supervised classification of down-hole physical properties measurements using neural network to predict the lithology. *Journal of Applied Geophysics*, **124**, pp.17-26.

Maji, S., Berg, A.C. and Malik, J. (2013). Efficient classification for additive kernel SVMs. *IEEE transactions on pattern analysis and machine intelligence*, **35**(1), pp.66-77.

Mambane, P. W., Hein, K. A. A., Twemlow, S. G. and Manzi, S. S. D. (2011). Pseudotachylite in the south boundary fault at the Cooke shaft, Witwatersrand Basin, South Africa. *South African Journal of Geology*, **114**(2), pp.109-120.

Manzunzu, B., Midzi, V., Mangongolo, A. and Essrich, F. (2017). The aftershock sequence of the 5 August 2014 Orkney earthquake (ML 5.5), *South African Journal of Seismology*, **21**(6), pp.1323-1334.

Maxwell, K., Rajabi, M. and Esterle, J. (2019). Automated classification of metamorphosed coal from geophysical log data using supervised machine learning techniques. *International Journal of Coal Geology*, 214, pp. 10384.

McDowell, G. M., King, A., Lewis, R. E., Clayton, E.A. and Grau, J. A. (1998). In-site nickel assay by prompt gamma neutron activation wireline logging. In *SEG Technical Program Expanded Abstracts, Society of Exploration Geophysicists*, pp.772-775.

McGarr, A., Simpson, D., Seeber, L. and Lee, W. (2002). Case histories of induced and triggered seismicity. *International Geophysics Series*, **81**(A), pp.647-664.

Meier, D.L., Heinrich, C.A. and Watts, M.A. (2009). Mafic dikes displacing Witwatersrand gold reefs: Evidence against metamorphic-hydrothermal ore formation. *Geology*, **37**(7), pp.607-610.

Mennan, A. (2017). Well log interpretation and 3D reservoir property modeling of Maui-B field, Taranaki Bain, New Zealand. Master Thesis submitted to Missouri University of Science and Technology.

Merembayev, T., Kurmangaliyev, D., Bekbauov, B., Amanbek, Y. A. (2021). A comparison of machine learning algorithms in predicting lithofacies: case studies from Norway and Kazakhstan. *Energies*, **14**, pp. 1896.

Midzi, V., Zulu, B., Manzunzu, B., Mulabisana, T., Pule, T., Myendeki, S. and Gubela, W. (2015). Macroseismic survey of the ML5.5, 2014 Orkney earthquake. *Journal of Seismology*, **19** (3), pp. 741-751.

Min, X., Pengbo, Q. and Fengwei, Z. (2020). Research and application of logging lithology identification for igneous reservoirs based on deep learning. *Journal of Applied Geophysics*, **173**, pp.103929.

Miyamoto, T., Hirono, T., Yokoyama, Y., Kaneki, S., Yamamoto, Y., Ishikawa, T., Tsuchiyama, A., Katayama, I., Yabe, Y., Ziegler, M., Durrheim, R.J. and Ogasawara. H. (2022). Characteristics of fault rocks within the aftershock Cloud of the 2014 Orkney Earthquake (M5. 5) Beneath the Moab Khotsong Gold Mine, South Africa. *Geophysical Research Letters*, **49**(14), p.e2022GL098745.

Mngadi, S. B. (2015). Integrated petrographic, geomechanical and seismological studies of rockmass behavior during the extraction of the Cooke 4 shaft pillar. A Dissertation submitted to the Faculty of Science, University of the Witwatersrand, Johannesburg.

Mostaghel, B. (1999). Processing and interpretation of geophysical well logs from well KJ-32, Krafla geothermal field, NE-Iceland. United Nations University. (8), pp. 193-220.

Nazeer, A., Abbasi, A. A. and Solagi, S. H. (2016). Sedimentary facies interpretation of Gamma Ray (GR) logs as basic well logs in Central and Lower Indus Basin of Pakistan. *Geodesy and Geodynamics*, **7**(6), pp. 432-443.

Nkosi, N.Z., Manzi, M.S., Westgate, M., Roberts, D., Durrheim, R.J., Ogasawara, H., Ziegler, M., Rickenbacher, M., Liebenberg, B., Onstott, T.C. and Team, D. (2022). Physical property studies to elucidate the source of seismic reflectivity within the ICDP DSeis seismogenic zone: Klerksdorp goldfield, South Africa. *International Journal of Rock Mechanics and Mining Sciences*, **155**, p.105082.

Nwaila, G.T., Zhang, S.E., Bourdeau, J.E., Negwangwatini, E., Rose, D.H., Burnett, M. and Ghorbani, Y. (2022). Data-Driven Predictive Modeling of Lithofacies and Fe In-Situ Grade in the Assen Fe Ore Deposit of the Transvaal Supergroup (South Africa) and Implications on the Genesis of Banded Iron Formations. *Natural Resources Research*, **31**(5), pp.2369-2395.

Ohakwere-Eze. M., Igboekwe. M., Chukwu. G. (2018). Petrophysical evaluation and lithology delineation using cross-plots analysis from some onshore wells in the Nigerian-delts, West Africa. *International Journal of Advanced Geosciences*, **6**(1), pp.99-107

Ogasawara, H., Durrheim, R. J., Yabe, Y., Ito, T., Aswegen, Van, G., Grobbelaar, M., Funato, A., Ishida, A., Mngadi, S. B., Manzi, M. S. D. and Ziegler, M. (2017). Drilling into seismogenic zones of M2.0 – M5.5 earthquake from deep South African gold mines (DSEIS): establishment of research sites. *Proceedings of AfriRock ISRM International Symposium, The Southern Africa Institute of Mining and Metallurgy, Cape Town, South Africa*, pp.237-248.

Ogasawara, H., Noda, T., Manzi, M. S. D., Durrheim, R. J. and Ogasawara, H. (2018). A capable geological structure of the Orkney M5.5 earthquake of 5 August 2014 identified by re-processing seismic reflection data. 20th *European Geosciences Union General Assembly, Proceedings from the conference held 4-13 April in Vienna*, Austria, pp.5470.

Ogasawara, H., Liebenberg, B., Rickenbacher, M., Ziegler, M., Onstott, T., Durrheim, R. J., Manzi, M. S. D., Mngadi, S., Yabe, Y., Ogasawara, H., Kaneki, S., Cason, E., Vermeuren, Jan-G., Heerden, E., Wiersberg, T., Zimmer, M., Kujawa, C., Conze, R., Aswegen, G., Wechsler, N., Ward, A. K., Enslin, S., Tau, S., Buccio, S. and the DSEIS team. (2019). status report: Drilling into seismogenic zones of M2.0-5.5 earthquakes in South Africa gold mines (DSEIS project), In W Joughin (ed.). *Proceedings of the Ninth International Conference on Deep and High Stress Mining, The Southern Africa Institute of Mining and Metallurgy*, pp.375-384.

Ortegon, J., Ledesma-Alonso, R., Barbosa, R., Castillo, J. V. and Atoche, A. C. (2018). Material phase classification by means of Support Vector Machines. *Computational Materials Science*, **148**, pp.336-342.

Pan, S. Y., Hsieh, B. Z., Lu, M. T. and Lin, Z. S. (2008). Identification of stratigraphic formation interfaces using wavelet and Fourier transforms. *Computers & Geosciences*, **34**(1), pp.77-92.

Parmar, H., Bhanderi, S. and Shah, G. (2014). Sentiment mining of movie reviews using Random Forest with Tuned Hyperparameters. In *International Conference on Information Science*, pp.1-6. Kerala.

Pechnig, R., Delius, H. and Bartetzko, A. (2005). Effect of compositional variations on log responses of igneous and metamorphic rocks. II: acid and intermediate rocks. *Geological Society, London, Special Publications*, **240**(1), pp.279-300.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V. and Vanderplas, J. (2011). Scikit-learn: Machine learning in Python. *The Journal of Machine Learning Research*, **12**(10), pp.2825-2830.

Rabaute, A., Yven, B., Chelini, W. and Zamora, M. (2003). Subsurface geophysics of the Phlegrean Fields: New insights from downhole measurements. *Journal of Geophysical Research: Solid Earth*, **108**(B3).

Ren, X., Hou, J., Song, S., Liu, Y., Chen, D., Wang, X. and Dou, L. (2019). Lithology identification using well logs: A method by integrating artificial neural networks and sedimentary patterns. *Journal of Petroleum Science and Engineering*, **182**, pp.106336.

Rencher, A.C. (2005). A review of "Methods of multivariate analysis, Second edition". IIE Transactions, 37(11), pp. 1083-1085

Rezaie, J., Sotrom, J. and Smorgrav, E. (2012), June. Reducing the dimensionality of geophysical data in conjunction with seismic history matching (spe 153924). In *74th EAGE conference and exhibition incorporating EUROPEC 2012* (pp. cp-293). *European Association of Geoscientists & Engineers.*

Rickenbacher, M. (2018). Geological and geotechnical rock mass model derived from two deep DSEIS boreholes drilled at Moab Khotsong, South Africa. MSc's thesis, pp.1-104.

Robb, L. J. and Meyer, F. (1995). The Witwatersrand Basin, South Africa, Geological framework and mineralization processes: *Ore Geology Reviews*, **10** (2), pp. 67-94.

Rodriguez, J. D., Perez, A. and Lozano, J.A. (2009). Sensitivity analysis of k-fold cross validation in prediction error estimation. *IEEE transactions on pattern analysis and machine intelligence*, **32**(3), pp.569-575.

Rodriguez-Galiano, V., Sanchez-Castillo, M., Chica-Olmo, M., Chica-Rivas, M. (2015). Machine learning predictive models for mineral prospectivity: An evaluation of neural networks, random forest, regression trees and support vector machines. *Ore Geology Reviews*.**71**, pp. 804–818.

Salehi, S.M. and Honarvar, B. (2014). Automatic identification of formation lithology from well log data: a machine learning approach. *Journal of Petroleum Science Research*, **3**(2), pp.73-82.

Schlumberger (1989). Log interpretation principles/applications. Schlumberger Wireline & Testing, Texas.

Sebtosheikh, M. A., & Salehi, A. (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. *Journal of Petroleum Science and Engineering*, **134**, pp.143-149.

Sevastyanov, L.A. and Shchetinin, E.Y. (2020). On methods for improving the accuracy of multi-class classification on imbalanced data. *Information and Telecommunication Technologies and Mathematical Modeling of High-Tech Systems* (*ITTMM)*, **20**, pp.70-82.

Shearer, P. M. (1996). Transition zone velocity gradients and the 520-km discontinuity. *Journal of Geophysical Research: Solid Earth*, **101**(B2), pp.3053-3066.

Stefansson, V. and Steingrimsson, B. S. (1990). Geothermal logging I, an introduction to techniques and interpretation, 3rd edition. Orkustofnun, Reykjavik report, pp.117.

Sun, J., Li, Q., Chen, M., Ren, L., Huang, G., Li, C. and Zhang, Z. (2019). Optimization of models for a rapid identification of lithology while drilling-A win-win strategy based on machine learning. *Journal of Petroleum Science and Engineering*, **179**, pp.321-341.

Tiab, D. and Donaldson, E. C. (2015). Petrophysics: theory and practice of measuring reservoir rock and fluid transport properties. Fourth Edition, ISBN:978-0-12-803188-9, Gulf professional publishing.

Uzoegbo, H.C. and Li, K. (2002). Mine-induced seismic events and its effect on nearby settlements in South Africa. *Retrieved on*, **5**, pp.01-16.

Vapnik, V. (1995). The nature of statistical learning theory. Springer science & business media, New York, ISBN 0-387-98780-0.

Wang, X., Hou, J., Song, S., Wang, D., Gong, L., Ma, K., Liu, Y., Li, Y. and Yan, L. (2018). Combining pressure-controlled porosimetry and rate-controlled porosimetry to investigate the fractal characteristics of full-range pores in tight oil reservoirs. *Journal of Petroleum Science and Engineering*, **171**, pp.353-361.

Xie, Y., Zhu, C., Zhou, W., Li, Z., Liu, X. and Tu, M. (2018). Evaluation of machine learning methods for formation lithology identification: A comparison of tuning processes and model performances. *Journal of Petroleum Science and Engineering*, **160**, pp.182-193.

Xu, H.J., Jin, Z.M. and Ou, X.G. (2006). Lithology determination of rocks from CCSD 100-2000 m main hole by magnetic susceptibility and density using discriminant function analysis. *Earth Sci J China Univ Geosci*, **31**(4), pp.513-519.

Yabe, Y. Abe, S., Ito, T., Ishida, A., Sugimura, K., Kanematsu, M., Higashi, M., Tadokoro, R., Ogasawara, H., Funato, A., Kato, H., Watson, B., Mngadi, S., Durrheim, R., Hofmann, G. and Scheepers, L. (2019). Stress state in the upper margin of aftershock zone of the 2014 Orkney earthquake (M5.5), South Africa, measured from boring cores of ICDP-DSeis project. AGU Fall Meeting, San Francisco, CA, 9-13 December.

Zhang, C., Zhang, Z. and Chen, Y. (2014). Logging lithology identification of tight sandy conglomerate and its application. *Science&Technology Review*, **33(**25), pp.59-62.

Zhang, D.Q., Zou, N.N. and Jiang, Y. (2015). Logging identification method of volcanic rock lithology: A case study from volcanic rock in Junggar Basin. *Lithologic Reservoirs (in Chinese)*, **27**(1), pp.108-114.

Zhang, J., Liu, S., Li, J., Liu, L., Liu, H. and Sun, Z. (2017). Identification of sedimentary facies with well logs: an indirect approach with multinomial logistic regression and artificial neural network. *Arabian Journal of Geosciences*, **10**(11), pp.1-9.

Zhang, S.E., Nwaila, G.T., Bourdeau, J.E., Frimmel, H.E., Ghorbani, Y. and Elhabyan, R. (2021). Application of machine-learning algorithms to the stratigraphic correlation of Archean shale units based on lithogeochemistry. *The Journal of Geology*, **129**(6), pp.647-672.

Zhao, L., Yang, J., Li, P. and Zhang, L. (2014). Seasonal inundation monitoring and vegetation pattern mapping of the Erguna floodplain by means of a RADARSAT-2 fully polarimetric time series. *Remote Sensing of Environment*, **152,** pp.426-440.

Zhao, S., Zhou, J. and Yang, G. (2019). Averaging estimators for discrete choice by M-fold cross-validation. *Economics Letters*, **174**, pp.65-69.

Ziegler, M., Rickenbacher, M., Berset, N., Liebenberg, B., Ishida, A., Sugimura, K., Noda, T., Ogasawara, H., Mngadi, S., Durrheim, R., Ito, T., Funato, A. and the ICDP DSEIS team. (2018). Core and borehole logging and stress measurements of a 817 m long borehole drilled from 2.9 km depth toward the Moab Khotsong 2014 M5.5 aftershock zone (ICDP DSeis project in South Africa). Abstract EGU2018-10377, Vienna, 12 April.

Zimmerman, D. W. (1994). A note on the influences of outliers on parametric and nonparametric tests. *Journal of General Psychology*, **121**(4), pp.391-401.

Zoppis, I., Mauri, G. and Dondi, R. (2019). Kernel methods: support vector machines.

Zou, Y., Chen, Y. and Deng, H. (2021). Gradient Boosting Decision Tree for Lithology Identification with Well Logs: A Case Study of Zhaoxian Gold Deposit, Shandong Peninsula, China. *Natural Resources Research*, **30**(5), pp. 1-21

Appendix A

Statistical distribution of lithology in Borehole A

	Lower Roodepoort quartzite				Crown basalt				Crown diabase intrusion			
	V_p (m/s)	GR (API)	V_s (m/s)	RHOB (g/cm^3)	V_p (m/s)	GR (API)	V_s (m/s)	RHOB (g/cm^3)	V_p (m/s)	GR (API)	V_s (m/s)	RHOB (g/cm^3)
Data size	711	711	711	711	898	898	898	898	783	783	783	783
Mean	5877	129.74	3698.49	2.68	6140	90.50	3746.94	2.81	6498	57.48	389.19	2.96
Std	84.36	42.27	55.57	0.06	120.71	17.23	54.58	0.06	90.81	18.26	44.19	0.06
Min	5695	63.24	3537.74	2.58	6031	42.00	3654.08	2.71	6360	18.08	3815.34	2.85
25%	5835	98.58	3658.54	2.66	6088	78.00	3710.35	2.78	6473	44.75	3880.98	2.93
50%	5879	119.35	3699.59	2.68	6106	90.15	3740.65	2.80	6521	53.38	3906.25	2.96
75%	5929	146.10	3739.49	2.71	6126	102.77	3754.69	2.82	6549	66.51	3926.70	2.99
Max	6065	217.21	3856.04	2.78	6179	135.43	3811.94	2.88	6620	99.06	3979.31	3.07

	Babrosco quartzite				Babrosco diabase intrusion			
	V_p (m/s)	GR (API)	V_s (m/s)	RHOB (g/cm^3)	V_p (m/s)	GR (API)	V_s (m/s)	RHOB (g/cm^3)
Data size	3183	3183	3183	3183	1323	1323	1323	1323
Mean	5709	72.96	3771.85	2.70	6621	7.42	3973.25	3.07
Std	177.29	30.15	81.88	0.07	49.85	4.49	44.27	0.05
Min	5457	0.80	3575.69	2.60	6520	0.80	3895.60	2.96
25%	5632	48.83	3723.01	2.67	6598	48.83	3957.78	3.04
50%	5690	70.81	3774.53	2.69	6627	70.81	3984.59	3.07
75%	5751	92.56	3821.66	2.71	6652	92.56	4001.07	3.10
Max	5913	157.86	3969.30	2.78	6731	19.36	4048.58	3.19

Appendix B

Statistical distribution of lithology in Borehole B

	Upper Roodepoort siltstone			Upper Roodepoort diabase intrusion			Upper Roodepoort quartzite		
	GR (API)	V_s (m/s)	RHOB (g/cm^3)	GR (API)	V_s (m/s)	RHOB (g/cm^3)	GR (API)	V_s (m/s)	RHOB (g/cm^3)
Data size	1012	1012	1012	300	300	300	4721	4721	4721
Mean	34.20	3553.12	2.79	15.78	3895.70	2.97	56.45	3734.37	2.69
Std	6.24	97.60	0.06	4.83	132.77	0.08	41.21	72.83	0.04
Min	20.35	3286.59	2.65	4.92	3807.75	2.90	13.88	3558.44	2.60
25%	29.95	3487.43	2.74	12.20	3897.45	2.97	39.39	3692.17	2..67
50%	33.52	3563.51	2.78	15.25	3941.84	3.00	48.66	3742.51	2.69
75%	38.01	3624.21	2.84	19.12	3968.25	3.02	60.83	3782.15	2.71
Max	48.48	3794.27	2.94	29.34	4017.46	3.07	92.91	3915.09	2.78

Lower Roodepoort siltstone				Lower Roodepoort diabase intrusion			Lower Roodepoort quartzite		
	GR (API)	V_s (m/s)	RHOB (g/cm^3)	GR (API)	V_s (m/s)	RHOB (g/cm^3)	GR (API)	V_s (m/s)	RHOB (g/cm^3)
Data size	1059	1059	1059	741	741	741	1300	1300	1300
Mean	35.83	3658.06	2.79	16.53	3967.64	3.06	51.07	3733.12	2.73
Std	9.27	74.32	0.07	6.32	72.27	0.06	22.46	84.60	0.05
Min	12.26	3482.43	2.65	4.82	3743.14	2.91	13.84	3566.90	2.59
25%	29.75	3602.88	2.82	12.90	3915.43	3.02	37.91	3697.16	2.69
50%	35.12	3648.45	2.78	15.87	3972.45	3.06	46.25	3750.63	2.72
75%	42.15	3712.87	2.82	18.73	4030.45	3.10	57.88	3780.06	2.76
Max	60.37	3865.98	2.93	27.03	4094.63	3.19	87.67	3908.29	2.86

	Crown basalt			Crown diabase intrusion		
	GR (API)	V_s (m/s)	RHOB (g/cm^3)	GR (API)	V_s (m/s)	RHOB (g/cm^3)
Data size	200	200	200	1040	1040	1040
Mean	63.64	3777.48	2.80	15.79	3951.53	2.97
Std	7.85	40.91	0.03	11.46	70.92	0.06
Min	43.73	3713.48	2.73	2.03	3792.35	2.86
25%	57.33	3755.63	2.78	8.96	3907.78	2.94
50%	63.76	3772.00	2.80	12.38	3951.53	2.97
75%	69.32	3783.97	2.82	18.27	3985.83	3.00
Max	85.26	3825.25	2.87	31.99	4067.98	3.08