International Journal of Innovative Computing, Information and Control
Volume **16**, Number **3**, June **2020**

ICIC International ©2020 ISSN 1349-4198
pp. 1077–1091

MISSING WELL LOG DATA HANDLING IN COMPLEX LITHOLOGY PREDICTION: AN NIS APRIORI ALGORITHM APPROACH

Touhid Mohammad Hossain[1,*], Junzo Watada[1], Zhiwen Jian[2]
Hiroshi Sakai[2], Shokor Rahman[3] and Izzatdin Abdul Aziz[1]

[1]Department of Computer and Information Sciences
[3]Department of Fundamental and Applied Sciences
Faculty of Information and Fundamental Sciences
Universiti Teknologi PETRONAS
Seri Iskandar, Perak Darul Ridzuan 32610, Malaysia
*Corresponding author: touhidmhossain@gmail.com
{ junzo.watada; shokor103072 }@gmail.com; izzatdin@utp.edu.my

[2]Department of Basic Sciences
Faculty of Engineering
Kyushu Institute of Technology
1-1 Sensui-cho, Tobata-ku, Kitakyushu-shi, Fukuoka 804-8550, Japan
jianzhiwen1996@yahoo.co.jp; sakai@mns.kyutech.ac.jp

Received January 2020; revised April 2020

Abstract. *Lithology prediction is considered an essential requirement in the field of petroleum exploration. Since reservoirs consist of complex lithologies, predicting the lithology classes is gradually playing a pivotal role in the geosciences. During drilling operations the advancements of real time data recording have been so common in the petroleum industries in the past and majority of the logging data are recorded in real time process. However, sometimes the system encounters data loss or missing values while going through the logging procedures. Hence, the application of missing data estimation in automated lithology prediction is so essential. In this research a unique module is developed for classifying lithology from borehole log data consisting of incomplete log values by employing non-deterministic information systems apriori (NIS Apriori) algorithm. The unique characteristics of the proposed module are also presented in the paper. The research proposes certain and possible rules based on real data science semantics following the framework of NISs. By using the NIS Apriori algorithm it is proved that each rule τ is determined by analyzing only a pair of τ-dependent possible tables although each particular rule τ is a dependant on so many possible tables. However, one of the applications of the NIS Apriori algorithm is its prospect of the handling missing values. This research proposes a white-box novel architecture to deal with the well log missing values by using the NIS Apriori algorithm which provides the results in terms of rules to classify complex lithology efficiently.*

Keywords: Lithology classification, Well log, NIS Apriori algorithm, Rule extraction, Completeness, Missing data

1. **Introduction.** Lithology refers to the composition or type of rock in the Earth's subsurface. Prediction of lithology is globally considered as a challenging problem in the primary step of petroleum exploration. The lithological data-points of reservoirs are indispensable for several geological factors like sedimentation modeling, and stratigraphic correlation [1-4]. As a reason, how to find accurate patterns from lithological information by utilizing data has become an important issue in the fields of petroleum geoscience [5,6].

DOI: 10.24507/ijicic.16.03.1077

Conventionally, the prominent statistics techniques that are used to classify lithology are principal component analysis (PCA), discriminant analysis, clustering analysis, etc. Giniyatullin et al. [7] used a type of clustering method for diagnosing lithological heterogeneity of the parent material. Fisher discriminant analysis is also applied to predicting lithology and it is found that the Fisher discriminant analysis is more effective than the linear discriminant analysis for classifying lithology [8,9]. Using different statistics methods Paasche and Eberle [10] processed geo-physical data sets and compared their application effects.

Recently, neural networks (NNs) are considered to be one of the most robust methods to classify lithology since the networks are built flexibly depending on the problem structure [11,12]. Several researchers used back propagation (BP) three-layered network for predicting lithologies because it is simple and robust [13]. Since in NN the training must be done analytically, which is generally calculated by least square algorithm, it is important that the data quality is well maintained [14]. That means, the data-points of well logs must not include false or missing data-values. However, in reality where several wells are drilled the well logs such as density or sonic and logs are sometimes found absent or missing because of some borehole problems or cost considering issues [15]. The absent log portions make it challenging to extract the information to classify the lithology. Additionally NN and most of the available methods for lithology classification are black-box or complex in terms of computational descriptions. Hossain et al. proposed a rough set based white-box approach to identify electrofacies to correlate lithology [16,17] but missing values were not considered in their research. This research proposes a white-box novel architecture to deal with the missing values by using a newly proposed algorithm which provides the results in terms of rules to classify complex lithology effectively and efficiently.

The methodology is related to rule generation using *Apriori algorithm* [18]. This work is also very much related to granular computing [19,20], rough sets (RSs) [21-23], data mining [18,24], and information incompleteness [25-30]. The mathematical information retrieval framework has been described by Marek and Pawlak [31]. The methodology corresponds to the origin of table data analysis and RS theory. We followed the systems and denote them as *nondeterministic information systems* (NISs). Correspondingly it is named, a table of information completeness, *deterministic information systems* (DISs). *Rough set theory* has been proposed by Pawlak [21,23] that provides a mathematical framework of table data analysis and generates rules. Numerous approaches and models also proposed rule based systems. Greco et al. [32] proposed a model of *dominance-RS* and used RS as tables whose attribute values are sorted. Ziarko [33] proposed *variable precision RS models* as an extension to RS. Komorowski et al. [34] made a survey on the framework of RS. Yao [35] proposed extension of RS to *three-way decisions* using probabilistic RS, and studied RS in *multi-granular spaces.*

Information incompleteness being a real data science problem has huge potential to work on but there is lack of research on this. Kryszkiewicz [26] characterized rules in incomplete information systems and missing attribute values are taken into DISs. Nakata and Sakai [28] worked on rule generation using possible world semantics following Lipski's incomplete information. Stefanowski and Tsoukiàs [36] worked on the relationship between rough classification and incomplete information tables.

Different models and approaches are shown by the researchers for classifying the generated rules using lower and upper approximations. For decision support, Predki et al. [37] developed an *RS Data Explorer*, and Bazan and Szczuka [38] developed an *RS Exploration System*, that is employable to knowledge discovery, data exploration and classification support. Grzymala-Busse [25] proposed learning from examples based on RS (LERS) and also extended the LERS system with missing data into table data. Ding et al. [39] worked with imbalance datasets. Sahri, et al. proposed an efficient non-parametric iteration based imputation method for dissolved gas dataset [40]. Typically statistical methods are used to handle missing values by imputation [8]. We used the framework of NIS Apriori rule generation by Sakai et al. [41] where Lipski's incomplete information databases and Orlowska's NISs are used and the methodology is related to RS that uses equivalence classes but the explanation of upper and lower approximations is somehow dissimilar. The purposes of our research are as below.

1) The dissimilarities between RS rule generation (RSRG) and Apriori rule generation (APRG) are clarified.
2) The methodology of rule generation using the NIS Apriori algorithm is described.
3) From NIS a process to approximate the actual DIS is shown.
4) The research demonstrates that NIS Apriori rule generation algorithm is basically an extension to RS, granular computing and three-way decisions and it can deal with missing values.
5) Finally, as an application to the methodology, 10 different lithology classes have been classified using several well log features that contain a number of missing values.

The study is organized as follows. Section 2 contains the methodology where, the differences between RSRG and APRG are discussed, DISs rule generation is extended to NISs and NIS Apriori algorithm is discussed. In Section 3 the experimental steps to handle missing values and to predict lithology using NIS Apriori algorithm are described. In Section 4 the comparison study is provided. Section 5 contains results and dicussion and Section 6 concludes the paper.

2. **Methodology.**

In lithology prediction, well logging is considered as an integral part and that can provide a huge amount of data to analyze. However, it is quite certain for the log records to be missing due to many factors such as instrument failure, broken instruments, borehole conditions or data loss due to inappropriate storage and incomplete logging [40]. As a consequence, some of the logging intervals get missing resulting it difficult to extract knowledge from the logs to predict lithology. Therefore, a robust method to deal with the missing logs becomes a necessity.

This section contains the methodology of generating rules and imputing the missing well log values using NIS Apriori algorithm for predicting lithology. As a background, DISs Apriori algorithm is also discussed in this section.

2.1. **DISs rule generation.**

Deterministic information systems (DISs), DISs Rules [23, 25,28,31,34], RSRG and APRG are discussed in this section.

2.1.1. *DISs rules.* DIS ψ signifies a quadruplet $\psi = (OBJ, ATR, \{VL_K | K \in ATR\}, m)$ in which OBJ indicates a finite set where the elements are named as *objects*, ATR indicates a finite set of *attributes*, VL_K is a finite set of *attribute values*, and m is a mapping where, m: $OBJ \times ATR \to \cup_{K \in ATR} VL_K$. We define $D \in ATR$ as the *decision attribute* and CND as a subset of $ATR \backslash \{D\}$ where CND is the *condition attributes* set. In ψ, $[K, vl]$ ($K \in ATR$, $vl \in VL_K$) is called a *descriptor*, and the formula τ: $\wedge_{K \in CND}[K, vl_K] \Rightarrow [D, vl]$ ($vl_K \in VL_K$, $vl \in VL_D$) is called an *implication*.

Definition 2.1. [23, 24, 41] *Considering DIS ψ, a pair of given thresholds $0 < \alpha, \beta \leq 1.0$, an implication τ containing support and accuracy as shown in items (1) and (2) below are the candidate of a particular rule in ψ.*

1) $supp(\tau)$ $(= |eq(\wedge_{K \in CND}[K, vl_K] \wedge [D, vl])|/|OBJ|) \geq \alpha$,
2) $acc(\tau)$ $(= |eq(\wedge_{K \in CND}[K, vl_K] \wedge [D, vl])|/|eq(\wedge_{K \in CND}[K, vl_K]|)) \geq \beta$

*Here, eq(*) denotes an object set that satisfies formula *, and* $|M|$ *indicates the cardinality of the set M (where M is an object set). If* $|eq(\wedge_{K,CND}[K, vl_K])| = 0$*, we define* $supp(\tau) = 0$ *and* $acc(\tau) = 0$*.*

2.1.2. *DISs Apriori rule generation.* Agrawal and Srikant proposed the Apriori algorithm to handle transaction data [18] and currently it has become a renowned algorithm for data mining. The algorithm has two useful properties that are shown below.

(Property 1) In APRG, considering two threshold values α and β, each descriptor $[K, vl_K]$ satisfying $|eq([K, vl_K])|/|OBJ| < \alpha$ is disregarded since any implication τ including $[K, vl_K]$ fails to satisfy $supp(\tau) \geq \alpha$. Hence, it is adequate considering descriptors in $\{[K, vl_K] || eq([K, vl_K])|/|OBJ| \geq \alpha\}$. The property decreases the number of insincere implications. Increased values for α reduce the number of the obtainable rules.

(Property 2) Considering η_1: $[K, vl_K] \Rightarrow [D, vl]$ and η_2: $[B, vl_B] \Rightarrow [D, vl]$ satisfying $acc(\eta_1) < \beta$ and $acc(\eta_2) < \beta$, $acc(\eta_3) \geq \beta$ occurs for η_3: $[K, vl_K] \wedge [B, vl_B] \Rightarrow [D, vl]$.

Therefore, if $supp(\eta_1) \geq \alpha$ and $supp(\eta_2) \geq \alpha$, then it is essential to consider η_3. Hence, the condition part of a given rule τ: $\wedge_{K\in CND}[K, vl_K] \Rightarrow [D, vl]$ includes the descriptors in $\{[K, vl_K] | supp([K, vl_K] \Rightarrow [D, vl]) \geq \alpha\}$. In Algorithm 1 IMP_{i+1} is generated from IMP_i and this process follows the above properties.

Algorithm 1: Adjusted Apriori algorithm for DIS [42].

```
Input: DIS $\psi$, decision attribute $D$, threshold values $\alpha$ and $\beta$;
Output: $Rule(\psi)$;
$Rule(\psi) \leftarrow \{\}$; $i \leftarrow 1$;
create $IMP_i$, where each $\tau_{i,j} \in IMP_i$ satisfies $supp(\tau_{i,j}) \geq \alpha$;
while ($|IMP_i| \geq 1$) do
    Rest $\leftarrow \{\}$;
    forall $\tau_{i,j} \in \beta$ do
        if $acc(\tau_{i,j}) \geq \beta$ then
            add $\tau_{i,j}$ to $Rule(\psi)$;
        else
            add $\tau_{i,j}$ to Rest;
        end
    end
    $i \leftarrow i + 1$;
    Generate $IMP_i$ via Rest and Property (2) mentioned above where $\tau_{i,j} \in IMP_i$
      satisfies:
    (a) $supp(\tau_{i,j}) \geq \alpha$, and
    (b) $\tau_{i.j}$ is not a redundant implication in the implication of $Rule(\psi)$
end
return $Rule(\psi)$;
```

In IMP_i the subscript i is the number of descriptors in the condition part, and $|IMP_i|$ is the number of implications. In another way, IMP_1, IMP_2, IMP_3 are sets of implications that consist of one, two and three condition attributes respectively.

Let τ': $(\wedge_{K\in CND}[K, vl_K]) \wedge [B, vl_B] \Rightarrow [D, vl]$ be a redundant implication for τ: $\wedge_{K\in CND}$ $[K, vl_K] \Rightarrow [D, vl]$. Assuming τ a rule, we can also regard that τ' is corresponding to a rule for decreasing the amount of rules. Every rule with least condition part is our target.

Proposition 2.1. [42] *Algorithm 1 is complete and sound for the generated rules in DIS* ψ*. Thus,* $Rule(\psi) = \{\tau | supp(\tau) \geq \alpha \text{ and } acc(\tau) \geq \beta \text{ in } \psi\}$ *holds in Algorithm 1.*

Proof: (Soundness) If $acc(\tau) \geq \beta$ holds, every implication $\tau \in IMP_i$ that assures $supp(\tau) \geq \alpha$, and τ is included to $Rule(\psi)$. This means each $\tau \in Rule(\psi)$ satisfies $supp(\tau) \geq \alpha$ and $acc(\tau) \geq \beta$ in ψ.

(Completeness) All implications τ (excluding the redundant cases) are considered to be in IMP_i, and the $acc(\tau)$ value is evaluated in Algorithm 1. So, there exists no implication τ that satisfies $supp(\tau) \geq \alpha$, $acc(\tau) \geq \beta$ and $\tau \notin Rule(\psi)$.

2.1.3. *Discussion of RSRG and APRG.* The subsection discusses the dissimilarities between RSRG and APRG considering two important aspects.

(Aspect 1: Rules characteristics)

RSRG: Let $eq([D, vl])$ signify set X. Some equivalence classes cover set X, and some rules τ: $\wedge_{K \in CND}[K, valA] \Rightarrow [D, vl]$ are extracted. The extracted rules are for the descriptor $[D, vl]$ which basically define set X.

APRG: In APRG set X is not definite, and it is noticed that, $X = OBJ$ (the whole set of objects). Therefore, some rules with higher support and accuracy are acquired.

(Aspect 2: Rule generation approaches)

RSRG: The set CND is potentially not exceptional considering a covering of X and as a consequence a few rules can be missed.

APRG: By using Apriori algorithm the implication τ can be obtained (except redundant implications): If τ satisfies $supp(\tau) \geq \alpha$ and $acc(\tau) \geq \beta$. Hence, completeness is ensured for the generated rules. However, for relatively lower threshold value of α APRG is time consuming. Although APRG and RSRG are different to some extent, both are useful rule generation algorithms. In RSRG, majority of the researches signify on the pair of approximations (lower and upper) of X but the pair of approximations does not acquire the set of rules directly. Our methodology mainly uses APRG.

2.2. **NISs rule generation and handling of missing data.** DIS is a standard table, and NIS is a table whose attribute value is a set of values. In NISs, possible world semantics is employed and each missing value is considered as a set of possible values.

For example, m(John,age) = 25 (John's age is 25) in DIS and m(Tom,age) = 23, 24, 25 (Tom's age is one of 23, 24, of 25) in NIS. If we do not know the details of his age, we often consider a set of possible values. The former information is deterministic information, and the latter is non-deterministic information. In m(Tom,age), we have three possible cases. Generally, we replace each set with a value of the set, and then we have DIS from NIS. We say such a DIS is a derived DIS from NIS. The image of the derived DISs is in Figure 1(a). The $Q(\phi)$ expresses the set of all derived DISs form NIS ϕ. We see there is one actual derived DIS due to the definition of NIS. In such $Q(\phi)$, we consider two types of rules below:

1) If an implication τ is a rule in each $\psi \in Q(\phi)$, we can conclude that this τ is also a rule in the actual derived DIS. Thus, we say τ is a certain rule.

2) If an implication τ is a rule at least one $\psi \in Q(\phi)$, we can conclude that this τ may be a rule in the actual derived DIS. Thus, we say τ is a possible rule.

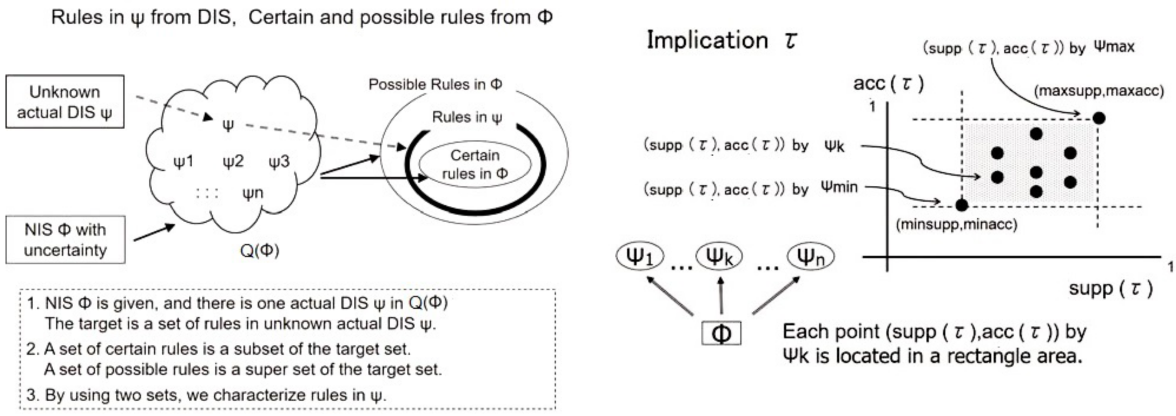
The definition of two types of rules seems natural in modal logic and possible world semantics. However, we face one problem. The quantity of $Q(\phi)$ increases exponentially. In the Mammographic data set in the UCI machine learning repository, the number exceeds 10 power 100. We cannot generate rules without some useful methods. We proposed one solution to handle this by Propositions 2.3 and 2.4 and Theorem 2.1. Based on the solution Algorithm 2 is developed.

Proposition 2.3 means that there is a derived DIS $\psi_{\min}$ where both *support* and *accuracy* are the minimum (denoted as $\min supp(\tau)$ and $\min acc(\tau)$ respectively in Definition 2.4). If τ is a rule in $\psi_{\min}$, τ satisfies below:

$supp(\tau)$ in $\psi \geq supp(\tau)$ in $\psi_{\min} \geq \alpha$,
$acc(\tau)$ in $\psi \geq acc(\tau)$ in $\psi_{\min} \geq \beta$.
Thus, τ is a rule in any ψ.

Proposition 2.4 means that there is a derived DIS $\psi_{\max}$ where both *support* and *accuracy* are the maximum (denoted as $\max supp(\tau)$ and $\max acc(\tau)$ respectively in Definition 2.4). If τ is a rule in $\psi_{\max}$, τ is a rule at least one derived DIS.

The intuitive meaning of them is expressed in Figure 1(b). Due to Propositions 2.3 and 2.4, we can conclude Theorem 2.1. Theorem 2.1 means that it is possible to examine certain rules and possible rules by checking $\psi_{\min}$ and $\psi_{\max}$.


Rules in ψ from DIS, Certain and possible rules from Φ
Unknown actual DIS ψ
ψ
ψ1 ψ2 ψ3
::: ψn
Possible Rules in Φ
Rules in ψ
Certain rules in Φ
NIS Φ with uncertainty
Q(Φ)
1. NIS Φ is given, and there is one actual DIS ψ in Q(Φ)
The target is a set of rules in unknown actual DIS ψ.
2. A set of certain rules is a subset of the target set.
A set of possible rules is a super set of the target set.
3. By using two sets, we characterize rules in ψ.
Implication τ
(supp (τ), acc (τ)) by Ψmax
acc (τ)
1
(maxsupp,maxacc)
(supp (τ), acc (τ)) by Ψk
(supp (τ), acc (τ)) by Ψmin
(minsupp,minacc)
Ψ1 ... Ψk ... Ψn
1
supp (τ)
Φ
Each point (supp (τ),acc (τ)) by Ψk is located in a rectangle area.

(a) Rules in the unknown actual DIS ψ and rules in NIS ϕ [44]

(b) $supp(\tau)$ and $acc(\tau)$ by $\psi \in \phi$ [42]

Figure 1. Visual representation of NIS Apriori algorithm

2.2.1. *NISs rules.* NIS ϕ is a quadruplet $(OBJ, ATR, \{VL_K | K \in ATR\}, m)$ where m denotes a mapping m: $OBJ \times ATR \rightarrow P(\cup_{K \in ATR} VL_K)$ (a power set of $\cup_{K \in ATR} VL_K$) [29,30]. Although the actual value is missing, each set $m(x, K)$ is interpreted as the actual value in the set. Here, information incompleteness is handled by mapping m.

Remark 2.1. *Considering NIS ϕ and $Q(\phi)$, let us accept that a DIS in $Q(\phi)$ contains actual information and let us name it as an unknown actual DIS ψ^{actual}. A method to select $\psi^{actual} \in Q(\phi)$ from ϕ is not present because of incompleteness of information. This concept of ψ^{actual} was presented for incomplete information databases [19] which we are following in our methodology. Now we can make a definition shown below.*

Definition 2.2. [42, 45] *Two types of new rules can be defined.*

1) A specific implication τ can be denoted as a certain rule, if τ denotes a rule in each $\psi \in Q(\phi)$.

2) A specific implication τ can be denoted as a possible rule, if τ denotes a rule in one $\psi \in Q(\phi)$ at least.

3) Let the set of certain rules and possible rules be $CR(\phi)$ and $PS(\phi)$ respectively.

2.2.2. *Certain and possible rules basics.* Basically, every certain rule is also considered as a possible rule. By observing NIS ϕ we can say in DIS ψ $m(x, K)$ is a singleton set for each x and K. Here, in this scenario $Q(\phi) = \{\psi\}$ holds which makes the definitions of certain and possible rules alike. Let us explain certain rules possible rules below.

Proposition 2.2. *Considering $\psi \in Q(\phi)$, the threshold values α and β, and Rule(ψ) in Algorithm 1, the below roles hold.*

1) $CR(\phi) = \cap_{\psi \in Q(\phi)} Rule(\psi)$,
2) $PS(\phi) = \cup_{\psi \in Q(\phi)} Rule(\psi)$,
3) $CR(\phi) \subseteq Rule\left(\psi^{actual}\right) \subseteq PS(\phi)$.

2.2.3. *Computational problem in NISs rule generation and the resolution.* As mentioned in Section 2.1.1, we experience the computational problem generating rule in NISs although Definition 2.2 seems well enough and in the rule generation process APRG is applied to each $\psi \in Q(\phi)$ though the number of the elements in $Q(\phi)$ exponentially increases. Therefore, it is difficult to sequentially apply APRG to each $\psi \in Q(\phi)$ and we have the computational-complexity issue for big data. We handled certain rules and possible rules in the following way to get rid of this problem and based on the resolution Algorithm 2 is proposed in Section 2.3.

Definition 2.3. [41, 45, 46] *Considering a mapping m in an NIS ϕ, two kinds of the granules named inf and sup are written as below.*
1) Considering a descriptor $[K, vl]$,
$inf([K, vl]) = \{x\colon object | m(x, K) = \{vl\}\}$,
$sup([K, vl]) = \{x\colon object | vl \in m(x, K)\}$.
2) Considering a conjunction $\wedge_{K \in CND}[K, vl_k]$ of descriptors,
$inf(\wedge_{K \in CND}[K, vl_K]) = \cap_{K \in CND} inf([K, vl_K])$.
$sup(\wedge_{K \in CND}[K, vl_K]) = \cap_{K \in CND} sup([K, vl_K])$.
If $m(x, K)$ indicates an item set for each x and K, we can observe that ϕ indicates a DIS and the same set is defined by the two kinds of granules named as inf and sup. An equivalence class in DIS is corresponded by the set. In NIS ϕ, all the equivalence classes are extended to inf and sup and the unknown actual equivalence class $eq([K, vl])$ fulfills $inf([K, vl]) \subseteq eq([K, vl]) \subseteq sup([K, vl])$.

Definition 2.4. [41, 45, 46] *Considering NIS ϕ and an implication τ, we can make some definitions as follows.*
1) $\min supp(\tau) = \min_{\psi \in Q(\phi)}\{supp(\tau) \in \psi\}$
2) $\min acc(\tau) = \min_{\psi \in Q(\phi)}\{acc(\tau) \in \psi\}$
3) $\max supp(\tau) = \max_{\psi \in Q(\phi)}\{supp(\tau) \in \psi\}$
4) $\max acc(\tau) = \max_{\psi \in Q(\phi)}\{acc(\tau) \in \psi\}$

Proposition 2.3. [41, 45, 46] *Considering NIS ϕ and an implication $\tau\colon \wedge_{K \in CND}[K, vl_K] \Rightarrow [D, vl]$, the following holds.*

$$\begin{aligned} &\min supp(\tau) = |inf(\wedge_{K \in CND}[K, vl_K]) \cap inf([D, vl])|/|OBJ|, \\ &\min acc(\tau) = \frac{|inf(\wedge_{K \in CND}[K, vl_K]) \cap inf([D, vl])|}{|inf(\wedge_{K \in CND}[K, vl_K])| + |OUTACC|} \\ &OUTACC = \{sup(\wedge_{K \in CND}[K, vl_K]) \backslash inf(\wedge_{K \in CND}[K, vl_K])\} \backslash inf([D, vl]) \end{aligned} \tag{1}$$

Furthermore, there is a defined DIS $\psi_{\min}$ where $supp(\tau) = \min supp(\tau)$ and $acc(\tau) = \min acc(\tau)$.

Proposition 2.4. [41, 45, 46] *Considering NIS ϕ and an implication $\tau\colon \wedge_{K \in CND}[K, vl_K] \Rightarrow [D, vl]$, the following holds.*

$$\begin{aligned} &\max supp(\tau) = |sup(\wedge_{K \in CND}[K, vl_K]) \cap sup([D, vl])|/|OBJ|, \\ &\max acc(\tau) = \frac{|sup(\wedge_{K \in CND}[K, vl_K]) \cap sup([D, vl])|}{|inf(\wedge_{K \in CND}[K, vl_K])| + |INACC|} \\ &INACC = \{sup(\wedge_{K \in CND}[K, vl_K]) \backslash inf(\wedge_{K \in CND}[K, vl_K])\} \cap inf([D, vl]) \end{aligned} \tag{2}$$

Furthermore, there is a defined DIS $\psi_{\max}$ *where* $supp(\tau) = \max supp(\tau)$ *and* $acc(\tau) = \max acc(\tau)$*. In Figure 1(b)* $supp(\tau)$ *and* $acc(\tau)$ *are shown by* $\psi \in \phi$*.*

Theorem 2.1. [41, 45] *Considering an implication* τ*, the following holds.*

1) τ *indicates a certain rule, iff* $\min supp(\tau) \geq \alpha$ *and* $\min acc(\tau) \geq \beta$*.*

2) τ *indicates a possible rule, iff* $\max supp(\tau) \geq \alpha$ *and* $\max acc(\tau) \geq \beta$*.*

3) In an NIS the process of generating rules by process (1) and (2) above is not dependent on the quantity of elements in $Q(\phi)$*.*

2.3. **NIS Apriori algorithm and rule generation.** In NIS Apriori based rule generation, we employ possible world semantics. The NIS Apriori algorithm (Algorithm 2) basically contains two phases, phase 1 is about generating the certain rule or $CRRule(\phi)$ and phase 2 is the generation of possible rule or $PSRule(\phi)$. Phase 1 uses the measured values $\min supp$ and $\min acc$ in Proposition 2.3, and phase 2 employs the measured values $\max supp$ and $\max acc$ as described in Proposition 2.4. However, as shown in Theorem 2.1, the number of derived DISs does not have effect on NIS Apriori algorithm. This gives us the indication that the NIS Apriori algorithm is appropriate to NIS with a huge number of derived DISs.

Algorithm 2: NIS Apriori algorithm [42].

```
Input: $NIS(\phi)$, the decision attribute $D$, the threshold values $\alpha$ and $\beta$.
Output: Two sets, $CRRule(\phi)$ and $PSRule(\phi)$
$CRRule(\phi) \leftarrow \{\}; i \leftarrow 1;$
create $IMP_i$, where each $\tau_{i,j} \in IMP_i$ satisfies $\min supp(\tau_{i,j}) \geq \alpha$
while ($|IMP_i| \geq 1$) do
    Rest $\leftarrow \{\}$;
    forall $\tau_{i,j} \in \beta$ do
        if $\min acc(\tau_{i,j}) \geq \beta$ then
            add $\tau_{i,j}$ to $CRRule(\phi)$;
        else
            add $\tau_{i,j}$ to Rest;
        end
    end
    $i \leftarrow i + 1;$
    Generate $IMP_i$ via Rest and Property (2) where $\tau_{i,j} \in IMP_i$ satisfies:
    (a) $\min supp(\tau_{i,j}) \geq \alpha$, and
    (b) $\tau_{i.j}$ is not a duplicate implication in the implication of $CRRule(\phi)$.
end
return $CRRule(\phi)$;
$PSRule(\phi) \leftarrow \{\}$;
$PSRule(\phi)$ generation follows the similar procedure of $CRRule(\phi)$ generation;
In this case, $CRRule(\phi)$, $\min supp(\tau_{i,j})$, and $\min acc(\tau_{i,j})$ are replaced $PSRule(\phi)$,
  $\max supp(\tau_{i,j})$, and $\min supp(\tau_{i,j})$.
return $PSRule(\phi)$;
```

Proposition 2.5. *The NIS Apriori algorithm is complete and sound for the defined certain rules* $CR(\phi)$ *and possible rules* $PS(\phi)$ *as in Definition 2.2 and they are equal to* $CRRule(\phi)$ *and* $PSRule(\phi)$ *in Algorithm 2, respectively.*

Proof: (Soundness: $CRRule(\phi) \subseteq CR(\phi)$ and $PSRule(\phi) \subseteq PS(\phi)$)

Each implication $\tau \in CRRule(\phi)$ satisfies $\min supp(\tau) \geq \alpha$ and $\min acc(\tau) \geq \beta$, so $\tau \in CR(\phi)$. The same is concluded for the possible rules as well.

(Completeness: $CR(\phi) \subseteq CRRule(\phi)$ and $PS(\phi) \subseteq PSRule(\phi)$)

In Algorithm 2, every implication τ is in IMP_i (except the redundant case as in Algorithm 1), and the condition of τ is validated. So, there exists no implication τ that satisfies the condition $\min supp(\tau) \geq \alpha$, $\min acc(\tau) \geq \beta$ and $\tau \in CRRule(\phi)$. The same is concluded for the possible rules.

3. **Lithology Classification Using NIS Apriori Algorithm.**

3.1. **Data acquisition.**

The purpose of the research includes the explanation of the special properties of the NIS Apriori rule generation algorithm, and the efficient use of it for handling missing logs in the process of lithology classification. However, we employed the dataset that consists of 10 well log attributes: Gamma Ray, Neutron Porosity, Density Correlation, Photoelectric Effect, Density Porosity, Conductivity, Caliper, Borehole Volume, Hole Diameter and Compressional Sonic. The unit, minimum, mean and maximum values of each well log attribute are shown in Table 1. We considered a total of 2741 tuples. From the geological core description dataset the lithological information is achieved which we considered as the decision attribute where 10 different lithology classes are found (as shown in Table 2).

In preprocessing step the dataset along with the lithology class information has been randomized and divided into two different datasets for training and testing and they are denoted as DTr and DTst respectively. DTr and DTst have 70 : 30 ratio which means DTr contains 1917 objects and DTst contains 834 objects. DTr contains 519 missing values in the dataset.

Table 1. Summaries of the selected well log attributes

Abbreviation	**GR**	**NPHI**	**DRHO**	**PE**	**DPHI**
Attribute	Gamma Ray	Neutron Porosity	Density Correlation	Photoelectric Effect	Density Porosity
Unit	.api	.decp	.g/cc	.none	.decp
Min.	46.49	0.1053	−0.05	0.5105	0.081
Mean	140.85	0.2942	0.02471	1.8684	0.254
Max.	385.89	1.2316	0.3244	3.7343	0.8082
Abbreviation	**CT10**	**CALI**	**BHVT**	**HDIA**	**DTC**
Attribute	Conductivity	Caliper	Borehole Volume	Hole Diameter	Compressional Sonic
Unit	.mmo/m	.in	.m3	.in	.uspf
Min.	113.9	5.297	0.4481	5.297	60.81
Mean	1240.4	5.863	2.0633	5.863	101.69
Max.	9215.5	7.58	4.265	7.58	159.61

Table 2. Classes and their corresponding lithologies

Class No.	**Lithology**	**Class No.**	**Lithology**
1	Claystone	6	Sandy Siltstone
2	Mudstone	7	Silty Sandstone
3	Siltston	8	Silty Mudstone
4	Sandstone	9	Muddy Sandstone
5	Sandy Mudstone	10	Granulestone

3.2. **Rule generation.** In this step, according to the methodology of NIS Apriori algorithm as described in Section 2.3 missing values in DTr are handled and certain rules $CRRule(\phi)$ and possible rules $PSRule(\phi)$ have been generated from dataset DTr in two steps as shown below.

Step 1: $CRRule(\phi)$ has been generated and the obtained rules satisfy $supp(\tau) \geq 0.005$, $acc(\tau) \geq 0.9$ for each possible table from DTr.

1) $CRRule(\phi_1)$, $CRRule(\phi_2)$, $CRRule(\phi_3)$, $CRRule(\phi_4)$ have been generated for 1 to 4 conditional attributes in this step.
2) It took less than one minute for generating $CRRule(\phi_1)$ to $CRRule(\phi_4)$.

Step 2: $PSRule(\phi)$ has been generated and the obtained rules satisfy $supp(\tau) \geq 0.00$, $acc(\tau) \geq 0.8$ for each possible table from DTr.

1) $PSRule(\phi_1)$, $PSRule(\phi_2)$, $PSRule(\phi_3)$, $PSRule(\phi_4)$ have been generated for 1 to 4 conditional attributes in this step.
2) It took less than one minute for generating $PSRule(\phi_1)$ to $PSRule(\phi_4)$.

Now, by applying the NIS Apriori algorithm on DTr, we have found that the certain rules have been generated to predict Claystone, Mudstone, Sandstone and Sandy Mudstone whereas the possible rules have been generated to predict the rest of the lithology classes. The details are shown in Table 3.

Table 3. Certain and possible rules and their corresponding lithologies

Certain rules set	**Lithology class**	**Possible rules set**	**Lithology class**
$CRRule(\phi_1)$	∅	$PSRule(\phi_1)$	∅
$CRRule(\phi_2)$	4	$PSRule(\phi_2)$	∅
$CRRule(\phi_3)$	1, 2, 4	$PSRule(\phi_3)$	6, 7, 8, 9, 10
$CRRule(\phi_4)$	1, 2, 4, 5	$PSRule(\phi_4)$	3, 6, 7, 8, 9, 10

3.3. **Estimation of lithology values in DTst.** We concluded the decision [Lithology,val] of each object from the applicable rule condition part ⇒ [Lithology, val] with the maximum lift value.

Table 4 shows the estimation results found from Step 1 and Step 2. The estimation process follows the steps mentioned below.

1) We applied each rule to each tuple in DTst.
2) We picked up one decision whose lift value is the highest.

$$lift = \frac{accuracy\ (P \Rightarrow Q)}{(Occurrence\ of\ Q/824)}.$$

Table 4. Estimation results

	Correctly estimated	**Incorrectly estimated**	**Nothing estimated**
Certain rules	527	86	211
Possible rules	12	10	189

From Table 4 we can calculate the lithology classification result. The numbers of total correctly and incorrectly estimated samples are 539 and 96. Total 189 samples have shown no result which we denoted as total nothing estimated. So,

Estimated ratio = 635/824 = 0.77,

Correctly estimated ratio = 539/635 = 0.85.

4. **Comparison Study.**

For comparison, two missing value imputation techniques and two prediction methods are taken into account.

4.1. **Imputation with different methods.**

In this step, the following methods are used for handling the missing values by imputation and dataset DTr is used for both of the imputation methods.

4.1.1. *Markov chain Monte Carlo method.* Markov chain Monte Carlo (MCMC) is a popular computer-driven sampling method for estimating posterior distributions in Bayesian inference [47,48]. It is a renowned method to handle the missing data [49,50]. MCMC method is used to impute the missing data in DTr.

4.1.2. *MissForest method.* MissForest is a random forest classifier (RFC) based imputation method where a random forest is trained depending on the observed values of a data matrix for making predictions on the missing values. It can be used to impute continuous and/or categorical data including complex interactions and non-linear relations [51]. RFC based imputation is also done to find the missing values in DTr.

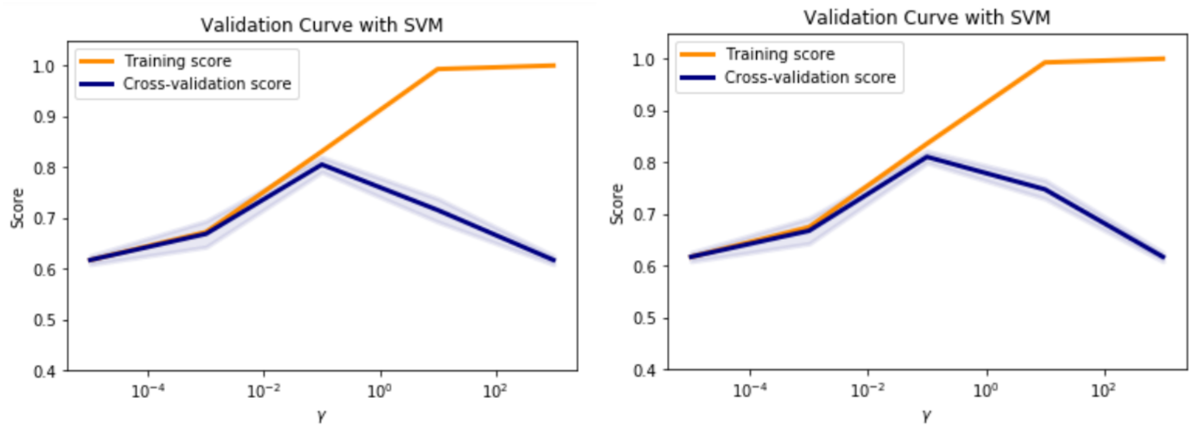
4.2. **Prediction accuracy for different methods.**

For both of the imputation methods prediction accuracy is calculated using support vector machine and random forest classifier.

4.2.1. *Support vector machine.* SVM is a machine learning tool proposed by Vladmir Vapnik in 1996 that has been used for 20 years to solve several problems, including lithology prediction [52,53]. In this step, SVM is used to calculate the lithology prediction accuracy. For this, the missing values in DTr are imputed using MCMC and MissForest to construct two different training sets and then the training sets are used to train the SVM module differently. Dataset DTst is used to validate the SVM module for both of the cases. The cross validation accuracy using SVM for both of the imputation methods are shown in Table 5. In Figure 2, the results are illustrated graphically.

Table 5. Cross validation scores for SVM and RFC

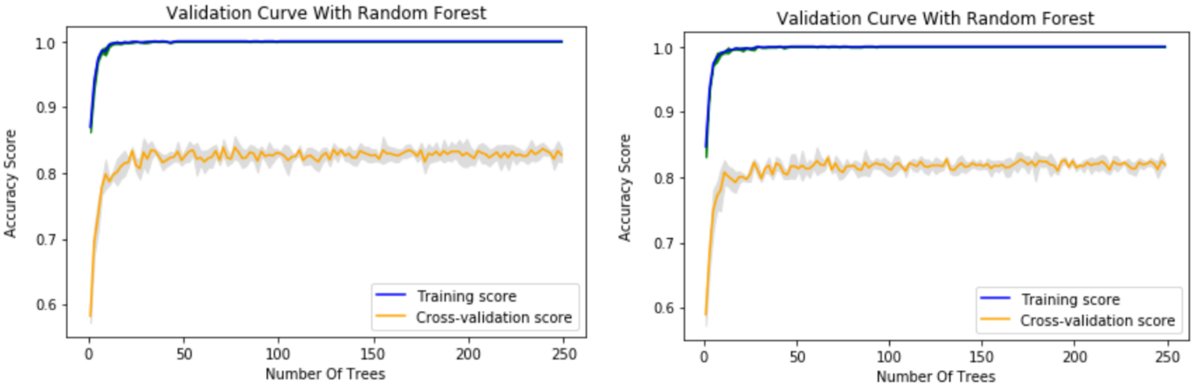
Imputation Method	**SVM**	**RFC**
MCMC	0.80461	0.83872
MissForest	0.81189	0.82312



(a) MCMC method (b) MissForest method

Figure 2. SVM scores for different imputation methods

4.2.2. *Random forest classifier.* Random forest classifier or RFC is an ensemble learning method for classification and regression that operates by constructing a multitude of decision trees while training and outputs the class based on mean prediction made by the individual decision trees. This method is used for solving several prediction problems including lithology prediction [50]. For comparison RFC is also used to calculate the prediction accuracy for dataset DTr after imputing the missing values by the two imputation methods MCMC and MissForest. We considered 250 decision trees to train the random forest and we validated the module using dataset DTst. The cross validation score using RFC for both of the imputation methods are shown in Table 5. In Figure 3, the results are illustrated graphically.



(a) MCMC method (b) MissForest method

FIGURE 3. RFC Scores for different imputation methods

5. **Results and Discussion.** In this paper a unique and efficient method is presented to deal with the missing well log values and to predict lithology using NIS Apriori algorithm. Out of 27470 log values there were 519 missing values and by using the methodology the missing values were handled and we have found that the lithology prediction accuracy is 0.85.

Finally, a comparison study is shown using two other renowned missing data imputation methods (MCMC and MissForest) and two other methods for classification (SVM and RFC) that have been widely applied in lithology prediction. Our findings indicate that, for lithology prediction the RFC performs slightly better than the SVM and the NIS Apriori method perfoms better than both SVM and RFC. Additionally, the easily explainable rules make the NIS Apriori module a white-box that can explain how the model produces the prediction of the lithology classes. By analyzing the rules, geoscientist can extract valuable information of how the features are contributing in forming a lithology class.

6. **Concluding Remarks.** In the paper, we discussed about the framework of DIS, Apriori and NISs and then studied the NIS Apriori algorithm on the basis of rule generation. For handling certain and possible rules by means of possible world semantics, the NIS Apriori algorithm is the only practically applicable algorithm. In the process of NIS Apriori rule generation, in Definition 2.2 $CR(\phi)$ and $PS(\phi)$ sets are discussed. These two sets named as lower and upper approximations of $Rule\left(\psi^{actual}\right)$ that we obtained in the unknown actual DIS ψ^{actual} are not similar to that of the typical RS. The concept of lower approximation $CR(\phi)$ and upper approximation $PS(\phi)$ of $Rule\left(\psi^{actual}\right)$ is similar to the concept of approximations in rough sets.

In the lithology classification dataset we found a number of missing log values and considering an application of NIS Apriori algorithm, we showed a unique method to handle ψ^{actual} which needs no further information.

From the results it is vivid that the NIS Apriori based rule generation is a unique method to deal with missing well log values in the process of classifying complex lithologies. Geophysicists and geologists can apply this methodology preferentially while needing well log datasets with missing values to complete other critical geology work. However, to guarantee a prediction module with better accuracy, an adequate training dataset is required. Choosing the appropriate explanatory features among a large number of well log attributes is also challenging. Furthermore, a larger dataset with more samples and datasets from different wells could also ensure a better prediction outcome. In future we intend to extend this module to calculate feature importance and to apply the methodology for predicting permeability and porosity from well log multi-variant dataset.

Acknowledgments. The authors would like to thank and highly appreciate Petroleum Research Fund (PRF), Cost Center 0153AB-A33 and the project leader Dr. Eswaran Padmanabhan for supporting the research. The authors like to thank Universiti Teknologi PETRONAS (UTP) for Graduate Research Assistantship (GRA) scheme.

REFERENCES

[1] A. Tomer, T. Muto and W. Kim, Autogenic hiatus in fluviodeltaic successions: Geometrical modeling and physical experiments, *J. Sediment. Res.*, vol.81, no.3, pp.207-217, 2011.

[2] P. M. Myrow, N. C. Hughes, M. P. Searle, C. M. Fanning, S. C. Peng and S. K. Parcha, Stratigraphic correlation of Cambrian-Ordovician deposits along the Himalaya: Implications for the age and nature of rocks in the Mount Everest region, *Geol. Soc. Am. Bull.*, vol.121, nos.3-4, pp.323-332, 2009.

[3] A. L. Larsen, M. Ulvmoen, H. Omre and A. Buland, Bayesian lithology/fluid prediction and simulation on the basis of a Markov-chain prior model, *Geophysics*, vol.71, no.5, pp.R69-R78, 2006.

[4] J. D. M. Travassos and P. D. T. L. Menezes, GPR exploration for groundwater in a crystalline rock terrain, *J. Appl. Geophys.*, vol.55, nos.3-4, pp.239-248, 2004.

[5] J. Jiang, Z. Rui, R. Hazlett and J. Lu, An integrated technical-economic model for evaluation CO_2 enhanced oil recovery development, *Appl. Energy*, vol.247, pp.191-211, 2019.

[6] Y. Chen, L. Lu and X. Li, Application of continuous restricted Boltzmann machine to identify multivariate geochemical anomaly, *J. Geochem. Explor.*, vol.140, pp.56-63, 2014.

[7] K. G. Giniyatullin, A. A. Valeeva and E. V. Smirnova, Application of cluster and discriminant analyses to diagnose lithological heterogeneity of the parent material according to its particle-size distribution, *Eurasian Soil Sci.*, vol.50, no.8, pp.917-924, 2017.

[8] S. Dong, Z. Wang and L. Zeng, Lithology identification using kernel Fisher discriminant analysis with well logs, *J. Pet. Sci. Eng.*, vol.143, pp.95-102, 2016.

[9] R. A. J. Aldrich, Fisher and the making of maximum likelihood 1912-1922, *Stat. Sci.*, vol.12, no.3, pp.162-176, 1997.

[10] H. Paasche and D. Eberle, Automated compilation of pseudo-lithology maps from geophysical data sets: A comparison of Gustafson-Kessel and fuzzy c-means cluster algorithms, *Explor. Geophys.*, vol.42, no.4, pp.275-285, 2011.

[11] M. Karmakar, S. Maiti, A. Singh, M. Ojha and B. S. Maity, Mapping of rock types using a joint approach by combining the multivariate statistics, self-organizing map and Bayesian neural networks: An example from IODP 323 site, *Mar. Geophys. Res.*, vol.19, pp.1-13, 2017.

[12] S. Sahoo and M. K. Jha, Pattern recognition in lithology classification: Modeling using neural networks, self-organizing maps and genetic algorithms, *Hydrogeol. J.*, vol.25, no.2, pp.311-330, 2017.

[13] B. Pradhan and S. Lee, Regional landslide susceptibility analysis using back-propagation neural network model at Cameron Highland, Malaysia, *Landslides*, vol.7, no.1, pp.13-30, 2010.

[14] C. W. M. Noor, R. Mamat and A. N. Ahmed, Comparative study of artificial neural network and mathematical model on marine diesel engine performance prediction, *International Journal of Innovative Computing, Information and Control*, vol.14, no.3, pp.959-969, 2018.

[15] Y. Gu, Z. Bao, X. Song, S. Patil and K. Ling, Complex lithology prediction using probabilistic neural network improved by continuous restricted Boltzmann machine and particle swarm optimization, *J. of Petroleum Sc. and Eng.*, vol.179, pp.966-978, 2019.

[16] T. M. Hossain, J. Watada, M. Hermana and H. Sakai, A rough set based rule induction approach to geoscience data, *International Conference of Unconventional Modelling, Simulation & Optimization on Soft Computing and Meta Heuristics (UMSO2018)*, Fukuoka, Japan, 2018.

[17] T. M. Hossain, J. Watada, M. Hermana and I. A. Aziz, Supervised machine learning in electrofacies classification: A rough set theory approach, *ICISE*, Kota Bharu, Malaysia, 2020.

[18] R. Agrawal and R. Srikant, Fast algorithms for mining association rules in large databases, in *Proc. of VLDB'94, Morgan Kaufmann*, J. B. Bocca, M. Jarke and C. Zaniolo (eds.), 1994.

[19] W. Pedrycz and A. Skowron, *Handbook of Granular Computing*, Wiley, 2008.

[20] L. A. Zadeh, Towards a theory of fuzzy information granulation and its centrality in human reasoning and fuzzy logic, *Fuzzy Sets Syst.*, vol.90, no.2, pp.111-127, 1997.

[21] Z. Pawlak, Rough sets, *Int. J. Comput. Inf. Sci.*, vol.11, no.5, pp.341-356, 1982.

[22] M. R. Salih, Y. F. Hassan and A. Elsayed, Data analysis using rough set theory and Q-learning algorithm, *ICIC Express Letters*, vol.13, no.4, pp.269-277, 2019.

[23] Z. Pawlak, *Rough Sets: Theoretical Aspects of Reasoning about Data*, Kluwer Academic Publishers, 1991.

[24] A. Skowron and C. Rauszer, The discernibility matrices and functions in information systems, in *Intelligent Decision Support – Handbook of Advances and Applications of the Rough Set Theory*, R. Słowiński (ed.), Kluwer Academic Publishers, 1992.

[25] J. W. Grzymala-Busse, Data with missing attribute values: Generalization of indiscernibility relation and rule induction, *Trans. Rough Sets I. Lecture Notes in Computer Science*, J. F. Peters, A. Skowron, J. W. Grzymała-Busse, B. Kostek, R. W. Świniarski and M. S. Szczuka (eds.), Berlin, Heidelberg, Springer, 2004.

[26] M. Kryszkiewicz, Rules in incomplete information systems, *Information Sciences*, vol.113, nos.3-4, pp.271-292, 1999.

[27] W. Lipski, On semantic issues connected with incomplete information databases, *ACM Trans. Database Syst.*, vol.4, no.3, pp.262-296, 1979.

[28] M. Nakata and H. Sakai, Twofold rough approximations under incomplete information, *Int. J. Gen. Syst.*, vol.42, no.6, pp.546-571, 2013.

[29] E. Orłowska and Z. Pawlak, Representation of nondeterministic information, *Theor. Comput. Sci.*, vol.29, nos.1-2, pp.27-39, 1984.

[30] Z. Pawlak, *Systemy Informacyjne: Podstawy Teoretyczne*, WNT Press, 1983 (in Polish).

[31] M. Marek and Z. Pawlak, Information storage and retrieval systems: Mathematical foundations, *Theor. Comput. Sci.*, vol.1, no.4, pp.331-354, 1976.

[32] S. Greco, B. Matarazzo and R. Słowiński, Granular computing and data mining for ordered data: The dominance-based rough set approach, in *Encyclopedia of Complexity and Systems Science*, R. A. Meyers (ed.), Springer, 2009.

[33] W. Ziarko, Variable precision rough set model, *J. Comput. Syst. Sci.*, vol.46, no.1, pp.39-59, 1993.

[34] J. Komorowski, Z. Pawlak, L. Polkowski and A. Skowron, Rough sets: A tutorial, in *Rough Fuzzy Hybridization: A New Method for Decision Making*, S. K. Pal and A. Skowron (eds.), Springer, 1999.

[35] Y. Yao, Three-way decisions with probabilistic rough sets, *Information Sciences*, vol.180, no.3, pp.314-353, 2010.

[36] J. Stefanowski and A. Tsoukiàs, Incomplete information tables and rough classification, *Comput. Intell.*, vol.17, no.3, pp.545-566, 2001.

[37] B. Predki, R. Słowiński, J. Stefanowski, R. Susmaga and S. Wilk, ROSE – Software implementation of the rough set theory, in *Rough Sets and Current Trends in Computing. Lecture Notes in Computer Science*, L. Polkowski and A. Skowron (eds.), Springer, 1998.

[38] J. G. Bazan and M. Szczuka, The rough set exploration system, in *Trans. Rough Sets III. Lecture Notes in Computer Science*, J. F. Peters and A. Skowron (eds.), Berlin, Heidelberg, Springer, 2005.

[39] L. Ding, J. Watada, L. C. Chew, Z. Ibrahim, L. W. Jau and M. Khalid, A SVM-RBF method for solving imbalanced data problem, *ICIC Express Letters*, vol.4, no.6(B), pp.2419-2424, 2010.

[40] Z. Sahri, R. Yusof and J. Watada, FINNIM: Iterative imputation of missing values in dissolved gas analysis dataset, *IEEE Trans. Ind. Inf.*, vol.10, no.4, pp.2093-2102, 2014.

[41] H. Sakai, R. Ishibashi, K. Koba and M. Nakata, Rules and Apriori algorithm in non-deterministic information systems, in *Trans. Rough Sets IX. Lecture Notes in Computer Science*, J. F. Peters, A. Skowron, H. Rybiński (eds.), Berlin, Heidelberg, Springer, 2008.

[42] H. Sakai, M. Nakata and J. Watada, NIS-Apriori-based rule generation with three-way decisions and its application system in SQL, *Information Sciences*, vol.507, pp.755-771, 2020.

[43] G. Biau, Analysis of a random forests model, *Journal of Machine Learning Research*, vol.13, pp.1063-1095, 2012.

[44] H. Sakai, K.-Y. Shen and M. Nakata, On two Apriori-based rule generators: Apriori in prolog and Apriori in SQL, *JACIII*, vol.22, no.30, pp.394-403, 2018.

[45] H. Sakai, M. Nakata and J. Watada, A proposal of machine learning by rule generation from tables with non-deterministic information and its prototype system, in *Proc. Int. Conf. on Rough Sets. Lecture Notes in Computer Science*, L. Polkowski et al. (eds.), Cham, Springer, 2017.

[46] H. Sakai, M. Wu and M. Nakata, Apriori-based rule generation in incomplete information databases and non-deterministic information systems, *Fundam. Inform.*, vol.130, no.3, pp.343-376, 2014.

[47] D. Gamerman and H. F. Lopes, Markov chain Monte Carlo: Stochastic simulation for Bayesian inference, in *Technometrics*, S. E. Ahmed (ed.), Boca Raton, Chapman and Hall/CRC, 2006.

[48] W. R. Gilks, S. Richardson and D. J. Spiegelhalter, *Markov Chain Monte Carlo in Practice*, Chapman and Hall/CRC, Boca Raton, 1996.

[49] L. Ingsrisawang and D. Potawee, Multiple imputation for missing data in repeated measurements using MCMC and copulas, *Proc. of the International MultiConference of Engineers and Computers Scientists*, Hong Kong, 2012.

[50] J. L. Schafer, *Analysis of Incomplete Multivariate Data*, Chapman and Hall, New York, 1997.

[51] D. J. Tekhoven and P. Buhlmann, MissForest – Non-parametric missing value imputation for mixed-type data, *Bioinformatics*, vol.28, no.1, pp.112-118, 2011.

[52] Wikipedia, *Support Vector Machine*, https://en.wikipedia.org/wiki/Support_vector_machine, Accessed on 11 April 2019.

[53] C. Deng, H. Pan, S. Fang, A. A. Konaté and R. Qin, Support vector machine as an alternative method for lithology classification of crystalline rocks, *J. Geophys. Eng.*, vol.14, pp.341-349, 2017.