processes

MDPI

Article

An Interpretable Predictive Model for Health Aspects of Solvents via Rough Set Theory

Wey Ying Hoo [1], Jecksin Ooi [1], Nishanth Gopalakrishnan Chemmangattuvalappil [2], Jia Wen Chong [2], Chun Hsion Lim [1] and Mario Richard Eden [3,*]

[1] School of Engineering and Physical Sciences, Heriot-Watt University Malaysia, No. 1, Jalan Venna P5/2, Precinct 5, Putrajaya 62200, Malaysia; wh25@hw.ac.uk (W.Y.H.); j.ooi@hw.ac.uk (J.O.); l.chun_hsion@hw.ac.uk (C.H.L.)
[2] Department of Chemical & Environmental Engineering, University of Nottingham Malaysia, Jalan Broga, Semenyih 43500, Malaysia; nishanth.c@nottingham.edu.my (N.G.C.); jiawen960218@gmail.com (J.W.C.)
[3] Department of Chemical Engineering, Auburn University, Auburn, AL 36849, USA
* Correspondence: edenmar@auburn.edu

Abstract: This paper presents a machine learning (ML) approach to predict the potential health issues of solvents by uncovering the hidden relationship between substances and toxicity. Solvent selection is a crucial step in industrial processes. However, prolonged exposure to solvents has been found to pose significant risks to human health. To mitigate these hazards, it is crucial to develop a predictive model for health performance by identifying the contributing factors to solvent toxicity. This research aims to develop a predictive model for health issues related to solvent toxicity. Among various algorithms in ML, Rough Set Machine Learning (RSML) was chosen for this work due to its interpretable nature of the generated models. The models have been developed through data collection on the toxicity of various organic solvents, the construction of predictive models with decision rules, and model verification. The results reveal correlations between solvent toxicity and the Balaban index, valence connectivity index, Wiener index, and boiling points. The generated predictive model using RSML has successfully provided insightful observations about the correlation between human toxicity and molecular attributes.

Keywords: machine learning; rough set theory; organic solvents; rough set-based machine learning; health indices

Citation: Hoo, W.Y.; Ooi, J.; Chemmangattuvalappil, N.G.; Chong, J.W.; Lim, C.H.; Eden, M.R. An Interpretable Predictive Model for Health Aspects of Solvents via Rough Set Theory. *Processes* **2023**, *11*, 2293. https://doi.org/10.3390/pr11082293

Academic Editor: Alexander Novikov

Received: 13 June 2023
Revised: 28 July 2023
Accepted: 28 July 2023
Published: 31 July 2023

1. Introduction

Solvents have been widely used in the chemical industry for dissolving, suspending, diluting, and separating substances. The paints and coatings sector holds the largest share in the global solvent market, followed by the printing inks segment, with the industrial cleaning industry in the third position [1]. However, prolonged exposure to solvents, particularly those containing volatile organic compounds (VOCs) can adversely affect human health, especially on respiratory, nervous, and reproductive systems. According to the National Institute of Occupational Safety and Health (NIOSH), approximately 9.8 million workers [2] are regularly exposed to high dosages of solvents each year through various exposure pathways. To address these health issues, organizations, such as the Occupational Safety and Health Administration (OSHA) and the U.S. Environmental Protection Agency (EPA), have established guidelines, such as permissible exposure limits (PEL), to detect and classify the associated health consequences.

While current research focuses on determining the potential issues associated with the use of organic solvents, recent research has highlighted the health issues related to solvent exposure [3]. The research findings revealed that the N-methyl-2-pyrrolidone (NMP) solvent, which is conventionally used to manufacture membranes, has neurotoxic, hepatoxic,

genotoxic, carcinogenic, and mutagenic effects on humans [3]. It is crucial to replace the toxic solvent with less toxic alternatives. Moreover, other researchers also reviewed the toxicokinetic, general toxicity, and reproductive toxicity associated with various frequently used solvents that primarily expose humans through inhalation routes [4]. The reports suggested that more detailed information on solvents is needed to understand the potential toxicity of solvents. However, the current research lacks a generalized predictive model for determining the health performance of solvents in a systematic approach. Although harmful substance concentration can be determined, the effect of toxicity can only be known after we apply a certain solvent into a process. Therefore, it is crucial to identify which molecular attributes will affect toxicity, whereby the results could then be applied in solvent design. This research aims to bridge the research gap between health issues and predictive models by determining the underlying relationship between solvents and health hazards based on their molecular structure.

This paper is divided into five main sections. The first section discusses the literature review on solvent toxicity, topological indices, and rough set machine learning (RSML), which led to the identification of research gap. Section 2 provides the detailed proposed methodology to close the identified research gap. Sections 3 and 4 mainly focus on the results and discussion by also highlighting the main insights gained from the obtained results. The summary and key contribution of this work, as well as potential future work, are illustrated in the last section.

1.1. Toxicity

Aromatic organic solvents represent 35% [5] of industrial utilization with high solvency [6] in forming solutions by dissolving a significant amount of solute. Based on the characteristic of high vapor pressure and low boiling point, solvents with a boiling point in the range of 50 °C to 260 °C [7] are known as volatile organic compounds (VOCs) and are easily emitted into the atmosphere at room temperature. The vaporized solvents can be readily absorbed by the human body through inhalation, ingestion, and skin, thereby affecting human health performance.

The impairment of health performance caused by exposure to VOCs varies depending on the exposure routes, levels, and type of solvents. The factors of exposure level, considering both concentration and duration, have been classified into short-term and long-term effects on human health, mainly through inhalation routes. The short-term health effects are dizziness, headache, nausea, and irritation in the eyes, nose, and throat [8]. The dispersed chemicals attach to the mucous layer of the membrane, leading to irritation and inflammation. In contrast, prolonged exposure to solvents has long-term health effects in terms of mutagenicity, toxicity, and carcinogenicity [8]. Thus, there is a necessity to investigate the long-term health effects of the solvents on the toxicity and carcinogenic parameter.

In the context of occupational health, the likelihood of solvent exposure through inhalation is significantly greater in terms of both quantity and frequency when compared to exposure through the oral and dermal routes [9]. Depending on the dosage, the ranking on the degree of toxicity can be classified according to Hodge and Sterner scale in lethal concentration 50% (LC50). Table 1 summarizes the toxicity rating by Hodge and Sterner Scale. Death caused by low dosage with less than or equal to 10 ppm indicates the substance would be classified as highly toxic as Class 1. Based on the rating of toxicity, organic solvents with a rating of 1 to 4 are considered toxic and must be avoided for extreme exposure.

Since the toxicity of organic solvents is related to their chemical structures, structural descriptors, such as Topological Indices (TI), have the potential to form a predictive model for toxicity. TIs are valuable tools in providing unique information about the structure of a molecule and are commonly used in predicting physicochemical properties of molecules.

Table 1. Toxicity rating based on Hodge and Sterner scale [10].

Toxicity Rating	Commonly Used Term	Inhalation LC50 (Exposure of Routes for 4 H) ppm
1	Extremely Toxic	10 or less
2	Highly Toxic	10–100
3	Moderately Toxic	100–1000
4	Slightly Toxic	1000–10,000
5	Practically Non-Toxic	10,000–100,000
6	Relatively Harmless	100,000

1.2. Topological Indices

Topological indices (TIs) are numerical values that characterize the structural features or properties of chemical compounds based on their molecular graphs [11]. The molecular topologies are influenced by the structure of the molecules in dimensions, configuration, bonding, symmetry, and degree of complexity. The most commonly used topological indices are the connectivity index, Wiener index, Randic index, Balaban index, and Zagreb index [12]. The TIs can be computed for any molecule based on their molecular structure. These indices can then be used to develop correlations in the studies of quantitative structure-activity (QSAR)/property (QSPR)/toxicity relationship (QSTR) [11].

TIs have been utilized for predicting toxicity and properties of chemicals [11]. For instance, the valence connectivity index, the Balaban index, and the electropy index were used to predict the underlying relationship between the molecular structure of ethers and toxicity in mice [13]. In toxicology, molecular topological indices have been applied to study the toxicity of alcohols, pesticides, and ionic liquids. In assessing the accuracy of the TIs to the toxicity, a figure of log LC50 versus TIs was plotted. A high accuracy and correlation will represent a quadratic model. In addition, the toxicity of organophosphorus pesticides is determined by the Randic–Kier–Hall connectivity indices and Topological Charge Indices (TCI) [14]. The indices are linked to the corresponding regression equation to calculate the LD50 value, and then, the calculated LD50 cross-validates with the experimental LD50 to identify the most correlated TIs. The studies have proven that the topological indices have an excellent correlation to the LD50. Besides, topological studies significantly reduce costs and save time compared to conducting experiments, making it an efficient approach [11]. However, there exists a research gap in connecting topological indices to a predictive model for assessing human health issues. With the incorporation of RSML, a predictive model can be developed based on topological indices of organic solvents.

1.3. Rough Set Machine Learning (RSML)

Machine learning (ML) is a specialized area of artificial intelligence (AI) that focuses on the automatic learning, analysis, and discovery of data [15]. Machine learning is capable of analyzing massive and fuzzy databases to discover patterns, reveal the underlying structure and relationships, and subsequently develop a robust prediction model. Based on learning methods, supervised learning in machine learning techniques determines the underlying data by imposing learning on the relationship between the past input-output training data through supervision. The supported algorithms are decision trees, support vector machines (SVM), K-Nearest Neighbors (KNN), RSML, Artificial Neural Network (ANN), and Bayesian Networks.

In a recent study, individual sets of mathematical tools, including fuzzy sets, rough sets, and soft sets, have been combined into a framework named Z-fuzzy soft β-covering-based rough matrices to solve a multiple attributes group decision-making (MAGDM) problem [16]. This combination has shown satisfactory results in recruiting the best applicant for the assistant professor job and can be further applied for decision-making problems. It further shows the capability and flexibility of rough sets to be combined with other tools in developing decision rules that would be useful for real-world problems.

RSML has its advantages in data processing without prior or complete information, whereas SVM approaches require maximum information in solving the problems of binary classification [17]. Moreover, RSML is preferred over ANN as RSML does not require human intervention and achieves similar accuracy within a shorter period [18]. Even though ANN have been utilized for decision-making tasks, such as Hepatitis B prediction and control in medical applications, there are challenges using ANN when it comes to uncertain data and uncommon diseases [19]. Although Frequent Pattern (FP) Growth algorithm is good in identifying patterns and generating association rules between the items based on support and confidence, it is unable to provide explanations understandable by humans and is less preferred when the data is uncertain and incomplete. Another commonly used machine learning method includes random forest (RF) due to its high predictive precision and flexibility. Nevertheless, there is a lack of interpretability that does not allow us to understand how a decision is made when compared to RSML. This is mainly because RF combines multiple decision trees, which makes it harder to interpret and understand the individual contributions of features. It becomes rather challenging to unveil meaningful physical insights through those aforementioned "black-box" ML approaches as relevance is often disclosed instead of focusing on the cause and effect [20]. The importance of having interpretable ML to drive knowledge generation has been emphasized in various research fields. For example, interpretable ML models are important in the electrocatalysis field to offer new insights into identifying novel catalytic materials and their mechanisms [20]. The recent advances of interpretable ML for estimating reactivity properties of solid surfaces and their existing challenges were also critically discussed in a recent contribution [21].

Due to the aforementioned benefits, RSML is capable of identifying and predicting the health performance of solvents by handling incomplete and uncertain data during important feature selection with excellent data efficiency and high versatility. Furthermore, RSML generates if-then decision rules in discovering the relationship between the conditional attributes of the objects to the decision attribute in a straightforward manner, then to be further applied to establish predictive models. In short, RSML benefits in data analysis and decision-making for datasets without requiring probability statistics and assumptions [22].

Rough set theory (RST) is a mathematical approach to analyzing imprecise and uncertain data or knowledge [23]. RST performs data classification, feature selection, and knowledge discovery tasks on large datasets. In theory, RSML functions are based on the approximation method. In the forms of approximation, the indiscernibility data exists within an elementary set bounded by lower and upper approximation. This is illustrated in Figure 1 [22]. The approximation method can be applied to toxic solvents, incorporating various pieces of information to perform classification based on the boundaries. With the use of lower and upper approximation to handle uncertainty caused by missing data, RSML is capable of handling uncertain and incomplete data. Likewise, RSML selects features by determining reducts, which are the minimal subsets of attributes affecting the decision attribute. This is important in generating meaningful decision rules that help in understanding and interpreting the data.

A rough set theory presents the information in a decision table consisting of an object, conditional attributes, and decision attributes. An example of an information table is presented in Table 2.

Table 2. Example of a decision table.

Object	Conditional Attributes		Decision Attribute
Type of Chemical	**Boiling Point (°C)**	**Wiener Index**	**Toxicity Class**
Toluene	100	2.5	Class 1

In the decision table, the object is the desired element for analysis, along with the conditional attributes [24] for the properties and characteristics of the corresponding object. The decision (class) attribute is determined based on the conditional input attributes

with generated decision rules [24]. By referring to Table 2, type of chemical is the object, conditional attributes comprise both boiling point and Wiener index, whereas toxicity class is the decision attribute. Both attributes can be quantitative (integer or decimal values) or nominal characteristics for numerical, categorical, and binary attributes. A structured information table in RSML benefits data classification, analysis, and discovery of the object-attribute relationship.

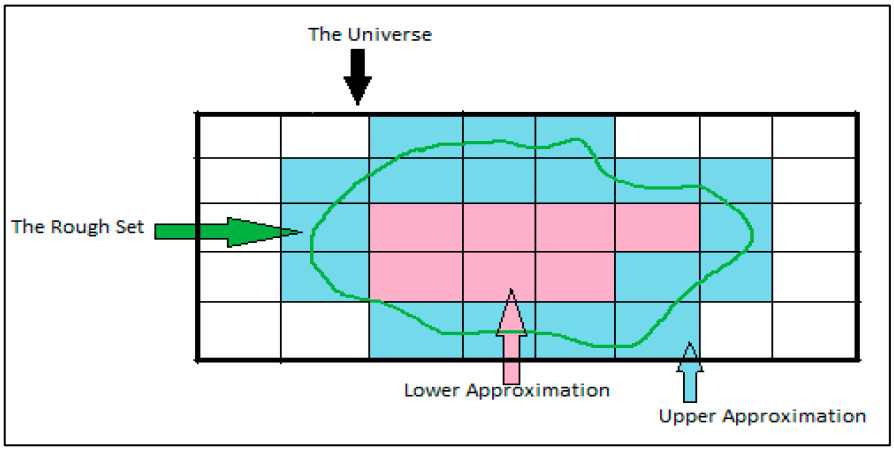

Figure 1. Illustration of Rough Set Theory.

RSML have been extensively applied for data mining and pattern recognition. For example, RSML integrates CO_2 capture and storage (CCS) for carbon management in selecting potential CO_2 storage sites [25]. RSML approaches are capable of performing a high certainty prediction on the storage sites, improving the CCS and negative emissions technologies (NETs) process. Besides storage sites, the RSML approach can also be applied to predict storage depth and geographical location. Moreover, the pyrolysis bio-oil properties can be predicted by RSML algorithms with the data on pyrolysis temperature and feedstock characteristics [26]. The focused pyrolysis bio-oil properties are the higher heating value and pH. The identified characteristics of feedstock samples are related to the amount of carbon, nitrogen, and oxygen content. In addition, another contribution was made by applying RSML in constructing a predictive model of the odor of fragrances [27]. RSML relates the odor properties in the form of topological indices and dilution as conditional attributes to the odor characteristics. Based on the rules induced, the fragrant topological indices in Kappa 3 and Kappa 2 dominate the characteristics. The fragrance prediction model through RSML greatly impacts the development of chemical products. Because of its interpretability, RSML was also used to construct predictive models for the estimation of physical and transport properties of polymers, such as glass transition temperature and cohesive energy [28]. The promising rules generated from RSML were then incorporated as property constraints in the computer-aided molecular design (CAMD) model to determine potential polymeric membrane molecular structure for air separation [28].

With the RSML applications above, rough sets theory with feature selection, clustering, and rule induction has promising outcomes in dealing with vague, ambiguous, and fuzzy datasets. As a result, the RSML algorithm is appropriate and applicable in developing an interpretable predictive model for determining the health performance of organic solvents. RSML efficiently aids in identifying the hidden structure and relationship with minimal resources and time consumption. Hence, this research aims to develop a predictive model of the health performance of organic solvents. The potential conditional attributes for the toxicity of organic solvents in human health focused on the topological indices and physical properties of solvents.

2. Methodology

This section explains the steps to identify the underlying structure and develop predictive models for health performance. The main steps include data collection (Step 1), developing a rough set model (Step 2), and verifying the prediction model (Step 3). The proposed methodology is illustrated in Figure 2.

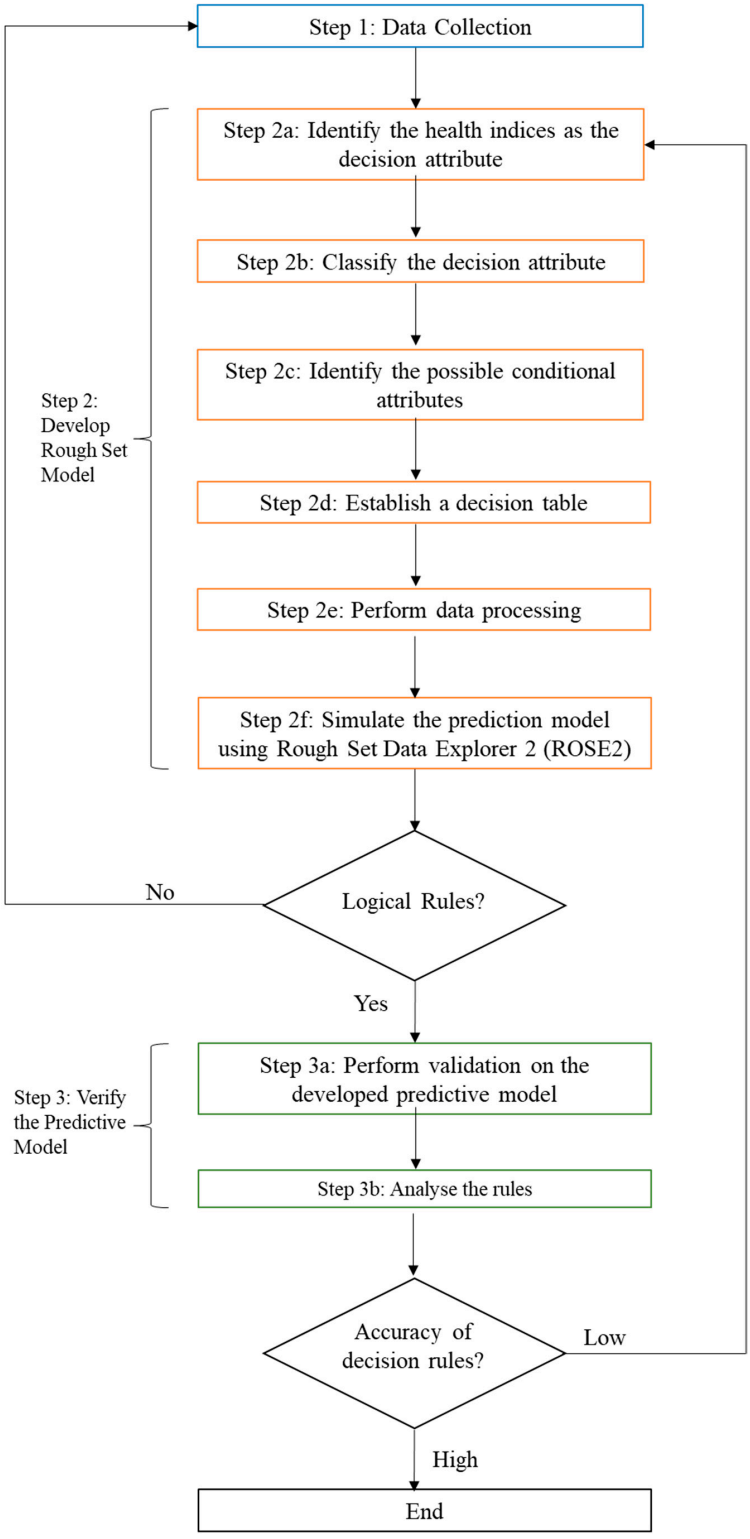

Figure 2. Proposed methodology to develop health performance predictive model.

Step 1 : Data collection on organic solvents

The first step involves collecting the toxicity data for a variety of chemicals which include aromatics, alcohols, ketones, ethers, esters, alkanes, and alkenes. The objects involved in this research are organic solvents for which 100 data points were collected from various sources and chemical compositions. Training data accounts for 70% of datasets for modeling, whereas 30% of datasets [26] preserve for validating the generated decision rules to ensure the accuracy of the predicted model. The database of organic solvents used for training and validation can be found in Appendix A Tables A1 and A2 respectively.

Step 2a : Identify the health indices as the decision attribute

Organic solvents cause various human health issues, such as toxicity and carcinogenicity. Therefore, this step identifies the targeted health issue of toxicity as a decision attribute in quantifying the toxic effects. The toxicity within this study encompasses all the toxic effects in the human body and is not restricted to particular human organs or systems.

Step 2b : Classify the decision attribute

Once the decision attribute has been identified, the attribute has to be classified and well-defined to distribute the object into the corresponding class. Based on the type of expression, the classification can be expressed in either category (categories 1,2,3), labels (high or low), or binary (yes/no) form. The toxicity as a decision attribute has been classified according to the Hodge and Sterner scale of standard LC50 through inhalation routes expressed in the form of categories (categories 1, 2, and 3).

In the predictive model, the toxic substances are classified into a simplified version based on Hodge and Sterner scale for rules induction, as shown in Table 3. For instance, the simplified toxicity rating 1 merges classes 1 and 2 from Hodge and Sterner scale. The toxicity classification for the predictive model of the health performance of solvents is then defined and classified into three main classes.

Table 3. Simplified toxicity rating for the prediction model.

Toxicity Rating	Commonly Used Term	Inhalation LC50 (Exposure of Routes for 4 H) ppm
1	Extremely to highly Toxic	<10–100
2	Slightly Toxic	100–100,000
3	Non-toxic	100,000

Step 2c : Identify the possible conditional attributes

A conditional attribute describes how the objects contribute to the decision attribute, expressed in a range or an exact value. The possible conditional attributes can be linked to the solvent's topological indices and physical properties.

In identifying conditional attributes, several topological indices related to human toxicity are chosen, including the Balaban Index, Wiener Index, and molecular connectivity index, whereas the chosen physical property is the boiling point of the solvent. The chosen topological indices have been supported by literature on the correlations of the Balaban index and valence connectivity index to toxicity [13].

Step 2d : Establish a decision table

The identified conditional attributes are then translated into quantifiable properties through calculations. For example, numerical values of relevant topological indices such as the Balaban index [29], the valence connectivity index [30], and the Wiener index [30] are calculated for each molecule. A decision table serves as a fundamental for the RSML model. Table 4 represents the simplified decision table for the organic solvent in the toxicity study.

Table 4. Simplified information table of organic solvent toxicity.

Object	Conditional Attributes				Decision Attribute
Organic Solvent	Balaban Index	Valence Connectivity Index	Wiener Index	Boiling Point (°C)	Toxicity Class
Acetyl Acetone	3.32	2.12	48	138	Class 1
1-Octanol	2.60	4.02	120	195	Class 2
Acetophenone	2.98	2.86	88	202	Class 3

From Table 4, the object represents the organic solvent, and the condition attributes comprise topological indices (Balaban Index, Valence Connectivity Index, Wiener Index) and the physical property (boiling point). The decision attribute in the toxicity classification is the three main classes based on the toxicity rating in Table 1. In rough set theory (RST), a decision table, also known as an information table made up of the universe, U (nonempty sets), and a set of attributes, A expressed in S = (U,A). The attribute comprised a set of values, V_a in the form of $a \in A$.

Step 2e : Perform data processing

The collected data must be pre-processed to eliminate redundant and unessential data before simulating the prediction model. In data processing, RST has the indiscernibility relation that describes the equivalence relation in a set of condition attributes, which denotes B⊆A [31]. In training for rule induction, the conditional attributes and respective classes undergo classification with approximation theory to indiscernibility. There are lower and upper approximations expressed in Equations (1) and (2) [31], respectively. The lower approximation is the set of attributes certain to belong to the class. In contrast, the upper approximation is the set of attributes that possibly fall in the subset.

$$B_*(X) = x \in U \{B(x) : B(x) \subseteq X\} \quad (1)$$

$$B^*(X) = x \in U \{B(x) : B(x) \cap X \neq \varnothing\} \quad (2)$$

The difference between the upper and lower approximation is named boundary region, expressed in Equation (3).

$$BN_B(X) = B^*(X) - B_*(X) \quad (3)$$

where $B(X)$ represents concept of decision X, U depicts the dataset, and BN_B represents boundary of the concept.

Furthermore, the data pre-processing process involves utilizing the concepts of reduct and core. In attribute reduction, reduct has the definition in determining the exact conditional attributes that impact the decision attribute, as shown in Equation (4) [32].

$$\gamma(A, X) = \gamma(A', X) \text{ for } A' \subseteq A \quad (4)$$

where A' represents a subset of the attribute set A and X is a decision class. A' is referred to as a reduct when the dependency of X on A' is identical to its dependency on A. It is worth noting that multiple reducts can exist for the same dataset. When this occurs, one can select reducts by considering criteria such as mechanistic plausibility or consistency with first principles. Whereas the core is defined as the identification of the intersection of all minimal subsets of attributes, as represented by Equation (5).

$$Core = \cap_{i=1}^{n} R_i \quad (5)$$

where R_i represents the ith reduct set.

Step 2f : Simulate the prediction model using Rough Set Data Explorer2 (ROSE2)

The generation of cores and reducts have been generated for model development via a software system, Rough Set Data Explorer, version 2 (ROSE2). ROSE2, based on rough set theory, has the core module coded with C++ programming [33] for the interface modules. ROSE2 performs data pre-processing with core and reduct functions and rules induction based on the Learning from Examples Module (LEM2) algorithm [33]. ROSE2 has a significantly lower 5% error rate than feature selection [34]. The generated rules from LEM2 are in the form of "IF-THEN" decision rules. The example of "IF-THEN" rule is as follows: "(Balaban Index in [34, 63)) & (Boiling Point $\geq$ 104), Decision: 1". The explanation for the example is if the molecule structure of an organic solvent has the Balaban index in the range of 34 to 63, along with a boiling point equal, and greater than 104 °C, then it will be classified as class 1 toxicity.

Step 3a : Perform validation on the developed predictive model

The developed predictive model with generated rules is validated with datasets from 30% of the collected data that are excluded from the training data. Validation is a vital step in rough set theory in evaluating the model and determining the appropriateness of the model for evaluating human toxicity from an organic solvent. The validation techniques numerically determine the strength, certainty, and coverage of models with Equations (4)–(6). High certainty and coverage indicate high satisfaction in the model performance with a determined underlying relationship. However, the low coverage and accuracy model can be re-modeled based on the previous steps.

Strength $\sigma_x(C, D)$ refers to the degree of supportiveness of the objects in the decision rule. Equation (6) shows the number of supported objects $supp_x(C, D)$ over the total available objects (U) in the decision table.

$$\sigma_x(C, D) = \frac{supp_x(C, D)}{|U|} \tag{6}$$

where C and D are the denotes for condition and decision attributes.

The certainty $(cer_x(C, D))$ shown in Equation (7) measures the probability of the conditional attribute being classified into the decision attribute. A high certainty percentage of 100% $(certainty\ factor,\ cer_x(C, D) = 1)$ indicates the generated rule is certain to the correct conditional attribute to be classified in the respective decision class, whereas the uncertain rule has the certainty factor to be less than 100% or a certainty factor of less than 1 $(0 < certainty\ factor,\ cer_x(C, D) < 1)$.

$$\begin{aligned} cer_x(C, D) &= \frac{|C(x) \cap D(x)|}{|C(x)|} = \frac{supp_x(C,D)}{|C(x)|} \\ &= \frac{\sigma_x(C,D)}{\sigma_x(C)} \end{aligned} \tag{7}$$

In simplified words, certainty is calculated with the object fulfilling the decision rule in the particular decision class $(\sigma_x(C, D))$ divided by all the objects that meet the decision rule $(\sigma_x(C))$.

Besides, coverage $(cov_x(C, D))$ measures the percentage of objects in the corresponding class under the rule.

$$\begin{aligned} cov_x(C, D) &= \frac{|C(x) \cap D(x)|}{|D(x)|} = \frac{supp_x(C,D)}{|C(x)|} \\ &= \frac{\sigma_x(C,D)}{\sigma_x(D)} \end{aligned} \tag{8}$$

Based on Equation (8), the coverage metrics determined by calculation with the object fulfilled the decision rule in the particular decision class $(\sigma_x(C, D))$ divide by the total number of objects in the particular decision class $(\sigma_x(D))$.

3. Results

3.1. Cores and Reducts

Five reduct sets were identified among four various conditional attributes of the Balaban index, valence connectivity index, Wiener index, and boiling point. The number of rules generated for each reduct is shown in Table 5. However, no core of the model was identified across reducts. The absence of a core indicates no constant conditional attribute intersects within the reduct sets. Thus, the results are analyzed based on the corresponding conditional attribute in each reduct.

Table 5. Reduct sets and the number of rules generated.

Reducts	Conditional Attributes	Number of Rules Generated
1	Balaban Index, Valence Connectivity Index	36
2	Valence Connectivity Index, Wiener Index	34
3	Balaban Index, Boiling Point	33
4	Valence Connectivity Index, Boiling Point	31
5	Wiener Index, Boiling Point	32

The study generated a sum of 166 decision rules within five sets of reduct. In reduct 1, 36 decision rules were generated, and they were all determined by the Balaban index and valence connectivity index. Reduct 2 was determined by the valence connectivity index and Wiener index, whereas reduct 3 has the Balaban index and boiling point. Furthermore, reduct 4 comprised valence connectivity index and boiling point. Lastly, reduct 5 consists of the Wiener index and boiling point. A summary of the reduct sets and the number of rules generated is shown in Table 5.

3.2. Rule-Based Prediction Models

In the predictive model, five different sets of decision rules, along with certainty and coverage, were generated. The complete set of decision rules from each reduct is shown in Appendix A Tables A3–A7. In order to perform rules interpretation, rules with good coverage and certainty are normally considered. Table 6 shows some examples of decision rules in reduct 5, and the rules will be used to illustrate the explanation.

Table 6. Decision rules in reduct 5.

No.	Rule	Decision	Coverage	Certainty
4	**(Balaban Index in [34, 63]) & (Boiling Point $\geq$ 104)**	**1**	32.26%	100%
22	**(Balaban Index in [17, 33]) & (Boiling Point in [36, 78])**	**2**	19.23%	100%
25	**(Balaban Index $\geq$ 153) & (Boiling Point $\geq$ 199)**	**3**	23.08%	100%

Rule 4 expressed in the "IF-THEN" statement, "If the molecular structure of organic solvent has the Balaban index in the range of 34 to 63, boiling point greater than or equal to 104 °C, then the inhalation LC50 falls within the range of 10 to 100 ppm, the organic solvent is classified as extremely to highly toxic in human health." The coverage of rule 4 was 32.26% by fulfilling 10 out of the 31 organic solvents of training data under class 1 toxicity. The certainty in the value of 100% indicated 10 organic solvents were classified in the correct decision attribute.

For class 2, the interpretation of the rules has the following "IF-THEN" statement "If the organic solvents have the Balaban index range of 17 to 33, boiling point between 36 °C to 78 °C, then the organic solvents were classified as class 2 toxicity as slightly toxic with inhalation LC50 in the range between 100 to 100,000 ppm." The coverage of rule 22 is comparatively lower than in class 1 in the value of 19.23% and a 100% certainty.

For class 3, the "IF-THEN" statement for the decision rule was "If the Balaban index of the organic solvents were greater or equals to 153, boiling point greater or equals to 199 °C, the organic solvents were classified under class 3 with relatively harmless to human health." The decision rule has 23.08% coverage and 100% certainty.

Similar to the remaining reduct sets, the interpretations for each decision rule are dominated by coverage and certainty. A high coverage indicates a high number of organic solvents in the particular class that met the rule's requirement. Meanwhile, certainty shows the accuracy of the objects classified under the correct class. Thus, the decision rule with high coverage and certainty obtains a rigid predictive model in predicting human toxicity in inhalation exposure to organic solvent.

3.3. Validation of Decision Rules

Thirty percent of datasets that was not utilized in developing the model were used to validate the decision rules. The validation data consist of exactly 30 data points of organic solvents applied to all five reduct sets in validating the coverage and certainty of the generated decision rules. The validation results exhibited high certainty for a class and were promising in showing the underlying relationship between the condition attributes and the decision attribute. Figures 3–7 demonstrate the validation results for each reduct set.

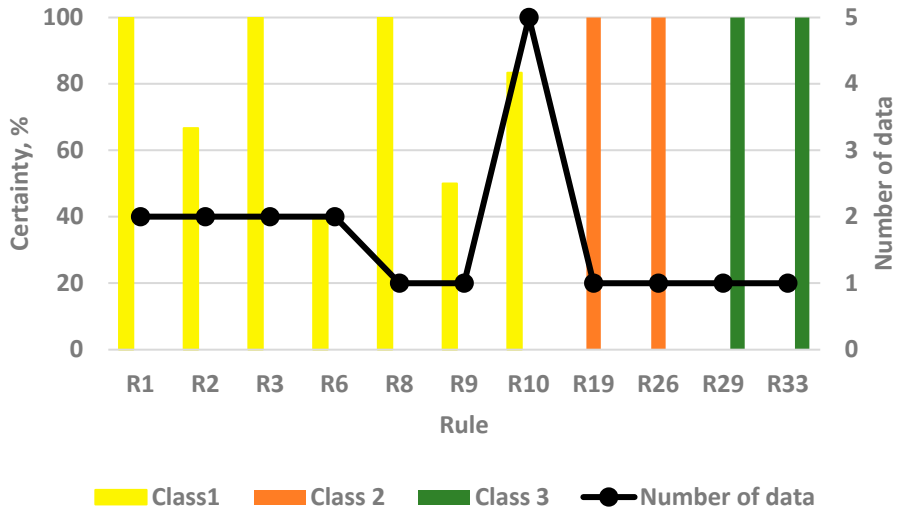

Figure 3. Validation results for reduct 1.

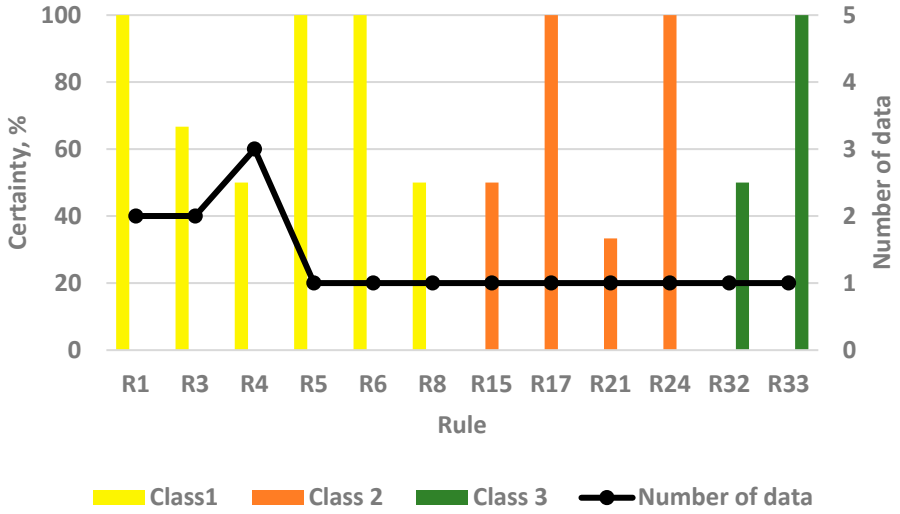

Figure 4. Validation results for reduct 2.

Figure 5. Validation results for reduct 3.

Figure 6. Validation results for reduct 4.

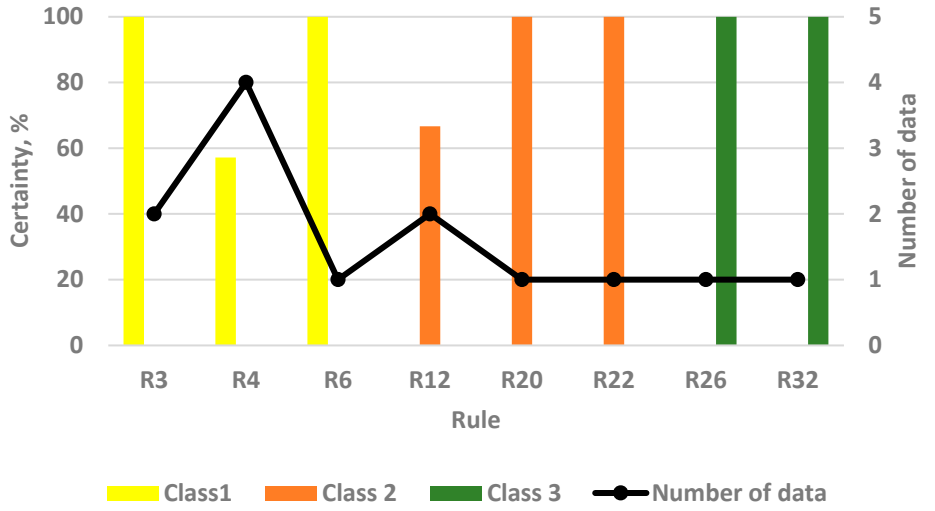

Figure 7. Validation results for reduct 5.

Reduct set has the minimal feature subset that correlates to the decision attributes in retaining the main backbone of the data set. For Class 1 toxicity, reduct 3 fulfilled the most

significant number of decision rules in 9 out of the 13 rules, followed by reduct 1 meeting 7 out of 15 rules in validation. In terms of certainty, Figure 5 illustrates that reduct 3 has 5 validated rules with 100% certainty, as compared to reduct 1, shown in Figure 3, which has only 4 rules with certainty higher than 70%. However, rule no.10 (R10) in reduct 1 meets a remarkable 5 data points in 83.33% certainty. Thus, reduct 1, with high to moderate certainty and the maximum coverage, was chosen for class 1 toxicity. In general, the conditional attributes of the Balaban index and valence connectivity index in reduct 1 affect the class 1 toxicity.

For class 2 toxicity, the validation datasets met four decision rules with R17 and R24 with 100% certainty. However, only a data point was present in each decision rule. Hence, in the review of reduct 5, with the second highest number of decision rules, the certainty for all three rules was comparatively high, with 66.67% certainty with 2 data points in R12 and 100% certainty with 1 data point for R20 and R22. With a high number of data points and certainty, reduct 5 correlates better to class 2 toxicity.

Molecules in class 3 were non-toxic to human health and were determined by reduct 3 with 3 decision rules. The Balaban index and boiling point mainly influence class 3 toxicity.

In summary, each toxicity class as a decision attribute is affected by various conditional attributes as shown in Table 7. The Balaban and valence connectivity indices from reduct 1 affect class 1 toxicity. Moreover, class 2 toxicity was influenced by the Wiener index and boiling point from reduct 5. Lastly, the Balaban index and boiling point from reduct 3 lead to the non-toxicity of class 3.

Table 7. Conditional attributes for each class.

Toxicity Classification	Conditional Attributes	Reduct
1	Balaban Index, Valence Connectivity Index	1
2	Wiener Index, Boiling Point	5
3	Balaban Index, Boiling Point	3

4. Discussion

Each toxicity class can be explained with conditional attributes based on the selected reduct set. The results in each class of decision attributes are interpretable with the chosen reduct set.

4.1. Class 1 Toxicity

Class 1 toxicity is defined as high to moderate toxicity of organic solvents via inhalation on the scale of LC50. The conditional attributes affecting class 1 toxicity include the Balaban index, denoted by A1, and the valence connectivity index, denoted by A2. In the validated decision rules, both topological indices (A1 and A2) have higher values in class 1 than in other classes. Table 8 shows the decision rules for classes 1 and 2 in reduct 1.

Table 8. Comparison of indices value among classes-reduct 1.

No.	Rule	Decision	Coverage	Certainty
8	**(A1 in [2.755, 2.89]) & (A2 $\geq$ 2.96)**	**1**	6.45%	100%
19	**(A1 in [2.78, 2.89]) & (A2 in [1.54, 2.55])**	**2**	11.11%	100%

The Balaban index, also known as the averaged distance sum connectivity [35], refers to the J index. J index increases with increasing branching and number of rings (aromatic). In other words, the J index was determined by molecules' branching (shape) in the proportional relationship. Besides, the valence connectivity index is the sum of overall bonds in counting the interacted bonding among two molecules in their valence states. This study focuses on the first-order valence connectivity index. The index measures the degree of connectivity between atoms based on the valence electron counts. The index correlates

with the polarity of a molecule in charge distribution. Then, the polarity is linked to the organic solvent's intermolecular forces and boiling point.

Solvent lipophilicity increases with molecular weight [36]. Hence, the increase of the Balaban index and aromaticity has increased the hydrophobicity of non-polar lipid-soluble molecules [37]. The dispersed organic solvent in the atmosphere with a high Balaban index and lipophilicity intake from inhalation tends to bind and accumulate in hydrophobic regions of the human body, like lipid-rich tissues. Moreover, the solvent with high valence connectivity index has a high degree of valence electron in uniform distribution resulting in low electronegativity and hence low polarity. The low polarity indicates weak intermolecular forces with low energy required for bond breaking. As a result, molecules with low polarity have a low boiling point. The low boiling point molecules tend to vaporize with high volatility. Therefore, the volatile solvent tends to disperse into the air and accumulate in the human body [35], risking extreme toxicity to human health. In summary, the topological indices of high Balaban index and valence connectivity index of the organic solvents are scientifically proven in high to moderate toxicity to human health via inhalation routes. Figure 8 summarizes the relationship of the topological indices in contributing towards high toxicity.

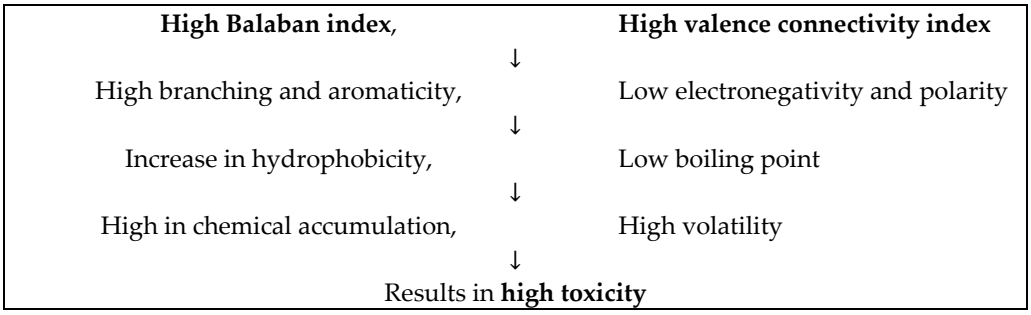

Figure 8. Summary of the effects of conditional attributes on class 1 toxicity.

In short, it can be concluded that organic solvents with high values of the Balaban index and valence connectivity index exhibit high to moderate toxicity to human health when inhaled.

4.2. Class 2 Toxicity

Class 2 toxicity is classified as slightly toxic to humans, with the inhalation routes of LC50 in rat exposure for 4 h ranging from 1000 to 100,000 ppm. Based on the validated decision rules, the conditional attributes of the Wiener index, denoted by A3, and the physical property of boiling points, denoted by A4, present in an organic solvent, are expected to contribute to class 2 toxicity. From the decision rules, class 2 is interpreted with a low Wiener index, with the lowest index lesser than 13 and a moderately high boiling point up to 288 °C.

In topological studies, the Wiener index quantifies the summation distances in the shortest path of each bonding [35] between two vertices. Wiener index correlates with molecular properties in QSAR and QSPR. The distance-based index has a good measurement of the compactness of a molecule [38] in an inversely proportional relationship. Hence, a molecule's Wiener index relates to its compactness and size. Moreover, boiling point is related to volatility. There is an inverse relationship between boiling point and volatility.

A slightly toxic organic solvent exhibits low Wiener index, indicating that the molecules are closely packed with large compactness. A compacted molecule was smaller, with vertices squeezed in a confined space. The small-sized molecules in the particle forms are easily inhaled into the lungs, then absorbed and distributed throughout the bloodstream. Simultaneously, organic solvents of volatile organic compounds (VOCs) in moderately high boiling points have a stronger intermolecular force, resulting in moderate volatility for molecules escaping into the atmosphere. Thus, small-sized particles with moderate volatil-

ity led to slight toxicity. A summary of the relationship between conditional attributes leading to toxicity is shown in Figure 9.

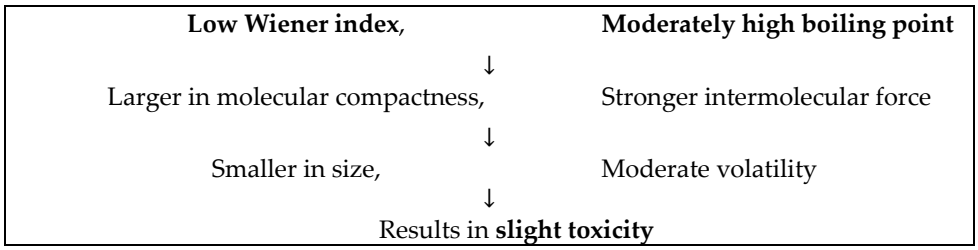

Figure 9. Summary of the effects of conditional attributes on class 2 toxicity.

Table 9 shows a few extracted decision rules from reduct 5 that result in class 2 toxicity. Rule 13 to 15 shows the binary relation of the Wiener index and boiling point on a decision rule. The rules of binary conditional attributes have the statement: "If the Wiener index is lower, then the solvent has a moderate boiling point", while in another case with the statement "If the Wiener index is high, the boiling point of solvent has to be higher, to be maintained in class 2 toxicity". On the other hand, the single conditional attribute of the Wiener index or boiling point can also contribute to toxicity. In rule 20, the Wiener index has a higher value in contributing towards slight toxicity, ranging from 72 to 83. The decision rule makes sense for having a higher value if a lower Wiener index of 10 in a single attribute would lead to highly toxic. Moreover, the conditional attribute of boiling point at 154 °C has been identified as class 2 toxicity.

Table 9. Extracted rules on class 2 toxicity from reduct 5.

No.	Rule
13	**(A3 $\geq$ 63) & (A4 in [189, 198])**
14	**(A3 < 10) & (A4 in [73, 83])**
15	**(A3 < 13) & (A4 in [117, 192])**
17	(A3 in [131, 153])
20	**(A3 in [72, 83])**
23	**(A4 = 154)**

Therefore, class 2 toxicity has characteristics of low Wiener index and moderately high boiling points of organic solvents. The compactness and intermolecular force of molecules allows for easy inhalation into the human body.

4.3. Class 3 Toxicity

The organic solvent with low toxicity is classified as class 3 with LC50 as 100,000 ppm in the inhalation routes. In the validation results, there is a high level of certainty for reduct 3 in class 3 toxicity. The conditional attributes, including the Balaban index denoted in A1 and the boiling point in A4, have been interpreted to dominate the organic solvent in meeting the criteria of class 3. A non-toxic effect of organic solvent is predicted to be in low Balaban index and high boiling point.

For an organic solvent with low toxicity, the solvents are low in the Balaban index, with lesser branching and aromaticity effects in decreasing lipophilicity. Low lipophilicity decreases the tendency of a molecule to be absorbed into body cells and tissues, resulting in low accumulation and hence lesser toxicity. In conjunction, a high boiling point solvent has a stronger intermolecular force and requires high energy in bond breaking, leading to less volatility. Therefore, solvents with the criteria of less complicated structure in low aromaticity and high volatility are relatively difficult in the phase change from liquid to a gas phase and interact with the human body, resulting in a non-toxic effect on the human body. An overall relationship between the Balaban index and boiling point is presented in Figure 10.

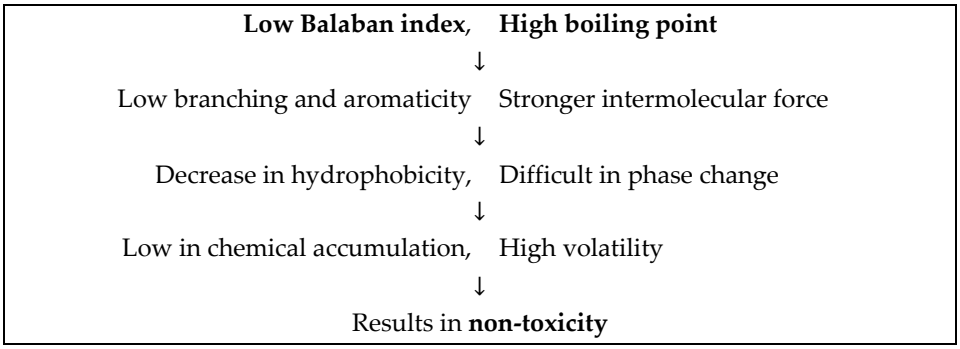

Figure 10. Summary of the effects of conditional attributes on class 3 toxicity.

The validated decision rules in Table 10 prove the attribute relationship between a low Balaban index and a high boiling point. In rule 27 and rule 28, the Balaban indices are lower than those in class 1, reduct 1. Similarly, a high boiling point is obtained in reduct 3 compared to class 3, reduct 5, tabulated in Table 10. The statement for rule 27 is explained as "If organic solvents with Balaban index ranging from 2.22 to 3.16, boiling point in the range of 199 °C to 284 °C, then the solvent is classified in class 3 toxicity. Moreover, a conditional attribute of a low Balaban index between 2.67 to 2.755 in rule 31 has resulted in extremely low toxicity. Hence, the validated decision rules for Class 3 toxicity highlight the significance of a low Balaban index and a high boiling point in determining the non-toxic effects of organic solvents.

Table 10. Extracted rules on class 3 toxicity from reduct 3.

No.	Rule
27	**(A1 in [2.22, 3.16]) & (A4 in [199, 284])**
28	**(A1 in [2.495, 2.625]) & (A4 in [159, 189])**
31	**(A1 in [2.67, 2.755])**

In summary, five reduct sets were determined from the data inputs in the predictive model. The generated decision rules have demonstrated validated data with reasonably good certainty and coverage and can be explained scientifically. When interpreting the results, the decision rules induced by the RSML approach showed the relationship between the molecular structure and the respective toxicity classification. When an organic solvent is classified as extremely toxic (class 1), it is mainly attributed to the high Balaban and valence connectivity indices. While a solvent is classified as slightly toxic (class 2), it was found to have low Wiener index and moderately high boiling point. Lastly, the low Balaban index and high boiling point are significant factors that contribute to low toxicity (class 3). As compared to other machine learning approaches, which are normally "black-box models" with limited explainability, this work has successfully revealed the key molecular attributes that lead to distinct classes of toxicity. This understanding holds significance in the process of designing new molecules or products.

5. Conclusions

This research paper presents the topological indices and physical properties of solvents as conditional attributes in a predictive model of human toxicity of solvents based on RSML. The impacts of solvents on human health can be estimated based on the factors, such as the boiling point of the solvent and the topological indices, including Balaban Index, the Valence Connectivity Index, and the Wiener Index. The predicted model based on uncertain and ambiguous data has generated rules to uncover the underlying structure of molecules contributing to human toxicity. The Balaban Index, valence connectivity index, and Wiener Index provide the quantitative values for the structural connection to the toxicity of organic solvents.

The proposed predictive model of the health performance of solvents with RSML has provided significant advantages to evaluate the health performance of solvents by discovering the conditional attributes that affect different classes of toxicity. This is particularly useful in solvent design and screening. However, the research has limitations on the assessment solely on human toxicity and may not account for other health issues caused by solvents. Further research should focus on developing prediction models for other health issues caused by solvents, such as carcinogenicity and mutagenicity. Additionally, future research directions should enhance the machine learning techniques and larger datasets of the predictive model with the incorporation of more comprehensive data in order to further improve the model's performance. In conclusion, this research successfully demonstrated the potential of using topological indices and physical properties of solvents in a predictive model using machine learning tools for assessing human toxicity.

Author Contributions: Conceptualization, J.O. and N.G.C.; Methodology, W.Y.H., J.O. and N.G.C.; Software, W.Y.H. and J.W.C.; Validation, J.O. and N.G.C.; formal analysis, W.Y.H.; Investigation, W.Y.H. and J.O.; resources, N.G.C. and J.W.C.; data curation, W.Y.H.; Writing—original draft preparation, W.Y.H.; Writing—review & editing, J.O., N.G.C., J.W.C., C.H.L. and M.R.E.; visualization, W.Y.H.; Supervision, J.O. and N.G.C.; project administration, J.O.; All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

Table A1. Database of organic solvents for training.

No.	Organic Chemical Solvents	Balaban Index	Valence Connectivity Index	Wiener Index	Boiling Point (°C)	Toxicity
1	2-Ethoxyethanol	2.34	2.10	35	135	1
2	Dimethylformamide (DMF)	2.83	1.39	18	153	1
3	Carbon tetrachloride	3.02	2.27	16	77	1
4	1-hexanol	2.45	3.02	56	156	1
5	Acetyl Acetone	3.32	2.12	48	138	1
6	Formamide	2.19	0.57	4	210	1
7	1,2-dichloroethane	1.97	2.10	10	84	1
8	1,2-dimethoxyethane (glyme, DME)	2.34	1.89	35	85	1
9	1,2-Dichloroethene	2.55	1.64	10	60	1
10	Nitromethane	2.80	0.81	9	101	1
11	Benzene	3.00	2.00	27	80	1
12	Trichloroethylene	3.14	2.08	18	87	1
13	Dibutyl ether	2.60	3.99	120	142	1
14	n-Decane	2.65	4.91	165	174	1
15	Methylisobutylketone	3.13	2.62	48	118	1
16	Pyridine	3.00	1.85	27	115	1
17	m-xylene	3.08	2.82	61	106	1
18	1,4-Dioxane	2.00	2.15	27	12	1
19	Methanol	1.00	0.45	1	65	1
20	Ethylene glycol	1.97	1.13	10	195	1

Table A1. *Cont.*

No.	Organic Chemical Solvents	Balaban Index	Valence Connectivity Index	Wiener Index	Boiling Point (°C)	Toxicity
21	Cyclohexanone	2.25	2.91	42	155	1
22	Ethylbenzene	2.83	2.97	64	136	1
23	Sulfolane	2.76	4.23	39	285	1
24	Chlorobenzene	3.02	2.48	42	132	1
25	Iso-Butanol	2.54	1.88	18	108	1
26	2-Methoxyethanol	2.19	1.51	20	125	1
27	Methylene chloride	1.63	1.60	4	40	1
28	p-xylene	3.03	2.82	62	106	1
29	o-xylene	3.13	2.83	60	106	1
30	Chloroform	2.32	1.96	9	61	1
31	1,1,2-Trichloroethene	2.54	2.52	18	114	1
32	1-propanol	1.97	1.52	10	97	2
33	2-Butanol	2.54	1.95	18	100	2
34	Pentene	2.40	2.02	20	30	2
35	Dimethyl Sulfoxide (DMSO)	2.80	2.95	9	189	2
36	Butyl acetate	2.82	2.90	79	125	2
37	Ethyl formate	2.40	1.47	20	55	2
38	2-Methyl-1-propanol	3.17	1.96	28	165	2
39	2-propanol	2.32	1.41	9	82	2
40	Pentane	2.19	2.41	20	36	2
41	Anisole	2.83	2.52	64	154	2
42	2-pentanone	2.83	2.76	32	102	2
43	Benzonitrile	3.05	2.38	64	190	2
44	Heptane	2.45	3.41	56	98	2
45	1-octanol	2.60	4.02	120	195	2
46	Methyl t-butyl ether (MTBE)	3.17	2.11	28	55	2
47	2-Amyl Alcohol	2.54	2.38	32	130	2
48	Hexamethylphosphoramide (HMPA)	4.69	5.03	142	231	2
49	Ethyl acetate	2.83	1.90	32	77	2
50	Formic acid	2.19	0.49	4	100	2
51	1-Butanol	2.19	2.02	20	118	2
52	Carbon disulfide	3.27	1.22	4	47	2
53	2-aminoethanol	1.97	1.22	10	171	2
54	Ethanol	1.63	1.02	4	79	2
55	Acetic Acid	2.80	0.93	9	118	2
56	Ethyl Ether	1.97	1.40	10	11	2
57	Diethylamine	2.19	2.12	20	56	2
58	Di-n-butyl phthalate	2.71	7.14	912	337	3
59	Glycerin	2.75	1.71	31	290	3
60	Ethyl benzoate	2.69	3.56	164	212	3
61	t-butyl alcohol	3.02	1.72	16	82	3
62	Ethyl acetoacetate	3.39	2.82	102	181	3
63	Dimethyl phthalate	3.15	3.96	295	283	3

Table A1. *Cont.*

No.	Organic Chemical Solvents	Balaban Index	Valence Connectivity Index	Wiener Index	Boiling Point (°C)	Toxicity
64	Diisopropyl ether	2.95	2.78	48	69	3
65	3-pentanol	2.75	2.49	31	115	3
66	Ether	2.19	1.99	20	35	3
67	Benzyl Alcohol	2.83	2.58	64	205	3
68	Diglyme (Diethylene glycol dimethyl ether)	2.60	2.97	120	162	3
69	Acetophenone	2.98	2.86	88	202	3
70	1,2-PropanedioI	2.54	1.56	18	188	3

Table A2. Database of organic solvents for validation.

No.	Organic Chemical Solvents	Balaban Index	Valence Connectivity Index	Wiener Index	Boiling Point (°C)	Toxicity
1	Toluene	3.02	2.41	42	111	1
2	Cyclohexanol	2.12	3.07	42	161	1
3	N,N-dimethylaniline	2.85	3.03	88	194	1
4	Tetrahydrofuran (THF)	2.08	2.08	15	65	1
5	Hexane	2.34	2.91	35	69	1
6	1,1-Dichloroethene	2.80	1.49	9	57	1
7	1-pentanol	2.34	2.52	35	138	1
8	Methylcyclohexane	2.12	3.39	42	101	1
9	N,N-Dimethylacetamide	3.26	1.82	29	165	1
10	Cyclohexane	2.00	3.00	27	81	1
11	1,1,1-Trichloroethane	3.02	2.20	16	75	1
12	n-octane	2.53	3.91	84	125	1
13	Acetonitrile	2.48	0.72	4	82	1
14	1-heptanol	2.53	3.52	84	176	1
15	N-methyl-e-pyrrolidone (NMP)	2.48	2.54	40	202	1
16	2-pentanol	2.63	2.45	32	119	1
17	Acetone	2.80	1.20	9	56	2
18	Aniline	3.02	2.20	42	184	2
19	Isopropyl acetate	3.13	2.30	48	89	2
20	Methyl acetate	2.85	1.32	18	57	2
21	Isobutyl acetate	3.05	2.76	74	117	2
22	Triethylamine	2.99	3.07	48	89	2
23	2-Butanone	2.85	1.76	18	80	2
24	1-Amyl Alcohol	2.34	2.52	35	138	2
25	Benzaldehyde	2.99	2.44	64	178	2
26	Diethyl ether	2.19	1.99	20	35	3
27	2,2,4-Trimethyl pentane	3.39	3.42	66	99	3
28	3-pentanone	2.99	2.33	31	102	3
29	Diethylene Glycol	2.45	2.21	56	246	3
30	Acetaldehyde	2.19	0.81	4	20	3

Table A3. Decision rules generated for reduct 1.

No	Rules		Toxicity Class	Training Data		Validation Data	
	Balaban Index	Valence Connectivity Index		Coverage (%)	Certainty (%)	Coverage (%)	Certainty (%)
1	2.495–2.625	3.215–4.005	1	3.23	100.00	12.50	100.00
2	2.22–2.37	≥1.58	1	12.90	100.00	12.50	66.67
3	<2.095	≥1.58	1	9.68	100.00	12.50	100.00
4	1.985–2.22	0.87–1.785	1	3.23	100.00	0.00	0.00
5	< 2.575	1.58–1.895	1	12.90	100.00	0.00	0.00
6	2.99–3.04	≥1.785	1	16.13	100.00	12.50	40.00
7	3.065–3.145	-	1	12.90	100.00	0.00	0.00
8	2.755–2.89	≥2.96	1	6.45	100.00	6.25	100.00
9	-	0.53–0.87	1	6.45	100.00	6.25	50.00
10	<2.545	2.445–3.215	1	9.68	100.00	31.25	83.33
11	-	1.305–1.395	1	3.23	100.00	0.00	0.00
12	<2.67	≥4.125	1	3.23	100.00	0.00	0.00
13	1.8–1.895	<1.175	1	3.23	100.00	0.00	0.00
14	3.295–3.355	-	1	3.23	100.00	0.00	0.00
15	<1.315	-	1	3.23	100.00	0.00	0.00
16	2.095–2.22	≥2.01	2	11.54	100.00	0.00	0.00
17	2.425–2.89	2.69–2.96	2	11.54	100.00	0.00	0.00
18	3.16–3.295	-	2	11.54	100.00	0.00	0.00
19	2.78–2.89	1.54–2.55	2	7.69	100.00	11.11	100.00
20	<1.985	1.175–1.54	2	11.54	100.00	0.00	0.00
21	2.37–2.625	1.895–2.445	2	11.54	100.00	0.00	0.00
22	-	0.87–1.075	2	7.69	100.00	0.00	0.00
23	<2.625	≥4.005	2	3.85	100.00	0.00	0.00
24	<2.495	≥3.215	2	3.85	100.00	0.00	0.00
25	-	1.395–1.49	2	11.54	100.00	0.00	0.00
26	3.04–3.065	-	2	3.85	100.00	11.11	100.00
27	≥4.04	-	2	3.85	100.00	0.00	0.00
28	≥1.315	<0.53	2	3.85	100.00	0.00	0.00
29	2.89–2.99	-	3	15.38	100.00	20.00	100.00
30	2.67–2.755	-	3	30.77	100.00	0.00	0.00
31	2.495–2.625	2.6–2.995	3	7.69	100.00	0.00	0.00
32	3.01–3.145	<1.995	3	7.69	100.00	0.00	0.00
33	-	1.975–1.995	3	7.69	100.00	20.00	100.00
34	3.145–4.04	≥2.485	3	15.38	100.00	0.00	0.00
35	-	2.55–2.6	3	7.69	100.00	0.00	0.00
36	-	1.54–1.58	3	7.69	100.00	0.00	0.00

Table A4. Decision rules generated for reduct 2.

No	Rules		Toxicity Class	Training Data		Validation Data	
	Valence Connectivity Index	Wiener Index		Coverage (%)	Certainty (%)	Coverage (%)	Certainty (%)
1	1.955–2.55	<19	1	16.13	100.00	12.50	100.00
2	1.58–1.675	-	1	6.45	100.00	0.00	0.00
3	2.8–3.215	34–63	1	16.13	100.00	12.50	66.67
4	<2.69	34–52	1	16.13	100.00	18.75	50.00
5	-	24–28	1	9.68	100.00	6.25	100.00
6	1.785–1.895	-	1	9.68	100.00	6.25	100.00
7	1.49–1.515	-	1	3.23	100.00	0.00	0.00
8	0.53–0.87	-	1	6.45	100.00	6.25	50.00
9	1.305–1.395	-	1	3.23	100.00	0.00	0.00
10	4.125–4.97	-	1	6.45	100.00	0.00	0.00
11	3.975–4.005	-	1	3.23	100.00	0.00	0.00
12	1.075–1.175	-	1	3.23	100.00	0.00	0.00
13	2.96–2.995	<71	1	3.23	100.00	0.00	0.00
14	<0.47	-	1	3.23	100.00	0.00	0.00
15	1.175–1.54	<13	2	19.23	100.00	11.11	50.00
16	2.01–2.135	19–29	2	15.38	100.00	0.00	0.00
17	2.69–2.77	-	2	3.85	100.00	11.11	100.00
18	≥3.215	52–72	2	3.85	100.00	0.00	0.00
19	1.895–1.975	≥17	2	11.54	100.00	0.00	0.00
20	≥4.005	63–153	2	7.69	100.00	0.00	0.00
21	2.325–2.445	-	2	11.54	100.00	11.11	33.33
22	0.87–1.075	-	2	7.69	100.00	0.00	0.00
23	2.88–2.905	-	2	3.85	100.00	0.00	0.00
24	<2.55	≥63	2	7.69	100.00	11.11	100.00
25	1.395–1.49	-	2	11.54	100.00	0.00	0.00
26	≥1.975	<10	2	3.85	100.00	0.00	0.00
27	0.47–0.53	-	2	3.85	100.00	0.00	0.00
28	2.77–2.8	-	3	7.69	100.00	0.00	0.00
29	<3.975	≥84	3	38.46	100.00	0.00	0.00
30	2.55–2.6	-	3	7.69	100.00	0.00	0.00
31	1.54–1.58	-	3	7.69	100.00	0.00	0.00
32	-	30–32	3	15.38	100.00	20.00	50.00
33	1.975–1.995	-	3	7.69	100.00	20.00	100.00
34	≥6.085	-	3	7.69	100.00	0.00	0.00
35	1.675–1.785	-	3	15.38	100.00	0.00	0.00

Table A5. Decision rules generated for reduct 3.

No	Rules		Toxicity Class	Training Data		Validation Data	
	Balaban Index	Boiling Point (°C)		Coverage (%)	Certainty (%)	Coverage (%)	Certainty (%)
1	<2.81	101–115	1	9.68	100.00	6.25	100.00
2	≥2.99	83–159	1	25.81	100.00	0.00	0.00
3	2.22–2.37	<67	1	3.23	100.00	6.25	20.00
4	1.985–2.37	≥121	1	12.90	100.00	12.50	66.67
5	-	131–154	1	19.35	100.00	6.25	50.00
6	2.425–2.495	≥101	1	3.23	100.00	6.25	50.00
7	≥2.99	58–81	1	6.45	100.00	6.25	100.00
8	<2.095	12–67	1	9.68	100.00	6.25	100.00
9	2.755–2.78	-	1	3.23	100.00	0.00	0.00
10	2.33–2.67	58–92	1	6.45	100.00	12.50	100.00
11	<1.985	≥173	1	3.23	100.00	0.00	0.00
12	2.625–2.67	-	1	3.23	100.00	6.25	100.00
13	<1.985	80–92	1	3.23	100.00	0.00	0.00
14	-	92–101	2	15.38	100.00	0.00	0.00
15	2.81–2.89	<128	2	11.54	100.00	22.22	100.00
16	≥2.625	154–172	2	7.69	100.00	0.00	0.00
17	2.545–2.625	≥172	2	3.85	100.00	0.00	0.00
18	-	189–192	2	7.69	100.00	0.00	0.00
19	-	164–172	2	7.69	100.00	0.00	0.00
20	≥2.22	<58	2	15.38	100.00	22.22	66.67
21	1.985–2.33	73–121	2	11.54	100.00	0.00	0.00
22	2.22–2.81	117–131	2	7.69	100.00	0.00	0.00
23	1.315–2.22	44–80	2	7.69	100.00	0.00	0.00
24	-	36–38	2	3.85	100.00	0.00	0.00
25	≥4.04	-	2	3.85	100.00	0.00	0.00
26	-	<12	2	3.85	100.00	0.00	0.00
27	2.22–3.16	199–284	3	30.77	100.00	20.00	50.00
28	2.495–2.625	159–189	3	15.38	100.00	0.00	0.00
29	3.01–3.025	81–116	3	7.69	100.00	0.00	0.00
30	-	67–73	3	7.69	100.00	0.00	0.00
31	2.67–2.755	-	3	30.77	100.00	0.00	0.00
32	3.355–4.04	-	3	7.69	100.00	20.00	100.00
33	-	33–36	3	7.69	100.00	20.00	100.00

Table A6. Decision rules generated for reduct 4.

No	Rules		Toxicity Class	Training Data		Validation Data	
	Valence Connectivity Index	Boiling Point (°C)		Coverage (%)	Certainty (%)	Coverage (%)	Certainty (%)
1	1.955–2.69	58–115	1	19.35	100.00	18.75	60.00
2	<2.135	121–159	1	12.90	100.00	0.00	0.00
3	2.55–2.69	<159	1	3.23	100.00	0.00	0.00
4	≥2.905	101–159	1	12.90	100.00	12.50	100.00
5	-	104–115	1	16.13	100.00	6.25	100.00
6	1.785–1.895	-	1	6.25	100.00	6.25	100.00
7	1.58–1.675	-	1	6.45	100.00	0.00	0.00
8	-	131–154	1	19.35	100.00	6.25	50.00
9	<1.175	≥121	1	6.45	100.00	0.00	0.00
10	4.125–4.97	-	1	6.45	100.00	0.00	0.00
11	0.53–0.87	-	1	6.45	100.00	6.25	50.00
12	≥1.49	<21	1	3.23	100.00	0.00	0.00
13	<0.47	-	1	3.23	100.00	0.00	0.00
14	1.895–2.55	≥154	1	11.54	100.00	22.22	50.00
15	≥2.01	21–58	2	15.38	100.00	0.00	0.00
16	≥2.05	189–198	2	11.54	100.00	0.00	0.00
17	0.47–1.54	<101	2	26.92	100.00	22.22	40.00
18	1.515–2.445	117–131	2	7.69	100.00	0.00	0.00
19	4.97–6.085	-	2	3.85	100.00	0.00	0.00
20	1.895–1.955	-	2	7.69	100.00	0.00	0.00
21	<1.305	104–172	2	7.69	100.00	0.00	0.00
22	≥2.88	<128	2	7.69	100.00	11.11	16.67
23	-	102–104	2	3.85	100.00	0.00	0.00
24	1.54–3.975	≥199	3	38.46	100.00	20.00	50.00
25	2.505–2.995	159–189	3	15.38	100.00	0.00	0.00
26	≥1.58	81–83	3	7.69	100.00	0.00	0.00
27	≥1.995	115	3	7.69	100.00	0.00	0.00
28	≥2.485	<73	3	7.69	100.00	0.00	0.00
29	1.975–1.995	-	3	7.69	100.00	20.00	100.00
30	≥6.085	-	3	7.69	100.00	0.00	0.00
31	-	178–189	3	15.38	100.00	0.00	0.00

Table A7. Decision rules generated for reduct 5.

No	Rules		Toxicity Class	Training Data		Validation Data	
	Wiener Index	Boiling Point (°C)		Coverage (%)	Certainty (%)	Coverage (%)	Certainty (%)
1	<24	101–115	1	9.68	100.00	0.00	0.00
2	-	83–92	1	9.68	100.00	0.00	0.00
3	<24	58–78	1	12.90	100.00	12.50	100.00

Table A7. *Cont.*

No	Rules		Toxicity Class	Training Data		Validation Data	
	Wiener Index	Boiling Point (°C)		Coverage (%)	Certainty (%)	Coverage (%)	Certainty (%)
4	34–63	≥104	1	32.26	100.00	25.00	57.14
5	<24	121–154	1	6.45	100.00	0.00	0.00
6	24–28	-	1	9.68	100.00	25.00	100.00
7	≥131	<198	1	3.23	100.00	0.00	0.00
8	-	131–154	1	19.35	100.00	0.00	0.00
9	<13	≥193	1	6.45	100.00	0.00	0.00
10	<6	<43	1	3.23	100.00	0.00	0.00
11	23–33	117–189	2	7.69	100.00	0.00	0.00
12	-	44–58	2	15.38	100.00	22.22	66.67
13	≥63	189–198	2	7.69	100.00	0.00	0.00
14	<10	73–83	2	7.69	100.00	0.00	0.00
15	<13	117–192	2	11.54	100.00	0.00	0.00
16	≥10	92–104	2	15.38	100.00	0.00	0.00
17	131–153	-	2	3.85	100.00	0.00	0.00
18	<32	117–121	2	7.69	100.00	0.00	0.00
19	-	92–101	2	15.38	100.00	0.00	0.00
20	72–83	-	2	3.85	100.00	11.11	100.00
21	<23	<32	2	7.69	100.00	0.00	0.00
22	17–33	36–78	2	19.23	100.00	11.11	100.00
23	-	154	2	3.85	100.00	0.00	0.00
24	≥30	115	3	7.69	100.00	0.00	0.00
25	≥153	≥199	3	23.08	100.00	0.00	0.00
26	45–131	≥199	3	15.38	100.00	20.00	100.00
27	≥13	81–83	3	7.69	100.00	0.00	0.00
28	-	159–163	3	7.69	100.00	0.00	0.00
29	≥30	<73	3	7.69	100.00	0.00	0.00
30	-	178–189	3	15.38	100.00	0.00	0.00
31	-	≥288	3	15.38	100.00	0.00	0.00
32	-	33–36	3	7.69	100.00	20.00	100.00

References

1. Future Business Insights. Industrial Solvents Market. In *Market Research Report*; Future Business Insights: Pune, India, 2019.
2. National Institute of Occupational Safety and Health. *Organic Solvent Neurotoxicity*; NIOSH Current Intelligence Bulletin 48. U.S. Dept. of Health and Human Services, Public Health Service, Centers for Disease Control, National Institute for Occupational Safety and Health: Cincinnati, OH, USA, 1987.
3. Tarrass, F.; Benjelloun, M. Health and environmental effects of the use of *N*-methyl-2-pyrrolidone as a solvent in the manufacture of hemodialysis membranes: A sustainable reflexion. *Nefrología (Engl. Ed.)* **2022**, *42*, 122–124. [CrossRef] [PubMed]
4. Vulimiri, S.V.; Pratt, M.M.; Kulkarni, S.; Beedanagari, S.; Mahadevan, B. Chapter 18—Reproductive and Developmental Toxicity of Solvents and Gases. In *Reproductive and Developmental Toxicology*, 3rd ed.; Gupta, R.C., Ed.; Academic Press: Cambridge, MA, USA, 2022; pp. 339–355. [CrossRef]
5. Ārija Baķe, M.; Eglite, M.; Martinsone, Ž.; Buiķe, I.; Piķe, A.; Sudmalis, P. Organic Solvents as Chemical Risk Factors of the Work Environment in Different Branches of Industry and Possible Impact of Solvents on Workers' Health. *Proc. Latv. Acad. Sci. Sect. B Nat. Exact Appl. Sci.* **2010**, *64*, 25–32. [CrossRef]

6. Stauffer, E.; Dolan, J.A.; Newman, R. Chapter 7—Flammable and Combustible Liquids. In *Fire Debris Analysis*; Stauffer, E., Dolan, J.A., Newman, R., Eds.; Academic Press: Burlington, NJ, USA, 2008; pp. 199–233. [CrossRef]
7. Soni, V.; Singh, P.; Shree, V.; Goel, V. Effects of VOCs on Human Health. In *Energy, Environment, and Sustainability*; Springer Nature: Singapore, 2018; pp. 119–142. [CrossRef]
8. Pruthu, K. Organic Solvents-Health Hazards. *J. Chem. Pharm. Sci.* **2014**, *3*, 83–86.
9. Institute of Medicine; Board on Health Promotion and Disease Prevention; Committee on Gulf War and Health: Literature Review of Pesticides and Solvents. *Gulf War and Health: Volume 2: Insecticides and Solvents*; National Academies Press: Washington, DC, USA, 2003; Volume 2.
10. Canadian Centre for Occupational Health and Safety, *What is a LD50 and LC50?* Canadian Centre for Occupational Health and Safety: Hamilton, ON, Canada, 2023.
11. Basak, S.C.; Mills, D.; Gute, B.D.; Grunwald, G.D.; Balaban, A.T. Applications of Topological Indices in the Property/Bioactivity/Toxicity Prediction of Chemicals. In *Topology in Chemistry*; Elsevier: Amsterdam, The Netherlands, 2002; pp. 113–184. [CrossRef]
12. Chemmangattuvalappil, N.G.; Eden, M.R. A Novel Methodology for Property-Based Molecular Design Using Multiple Topological Indices. *Ind. Eng. Chem. Res.* **2013**, *52*, 7090–7103. [CrossRef]
13. Bonchev, D. Applications of Topological Indices to QSAR. The Use of the Balaban Index and the Electropy Index for Correlations with Toxicity of Ethers on Mice. *Acta Pharm. Jugosl.* **1987**, *37*, 75–86.
14. García-Domenech, R.; Alarcon-Elbal, P.; Bolas, G.; Bueno-Marí, R.; Chordá-Olmos, F.A.; Delacour, S.A.; Mouriño, M.C.; Vidal, A.; Gálvez, J. Prediction of acute toxicity of organophosphorus pesticides using topological indices. *SAR QSAR Environ. Res.* **2007**, *18*, 745–755. [CrossRef]
15. Kononenko, I.; Kukar, M. Machine Learning and Data Mining. In *Machine Learning and Data Mining*; Elsevier: Amsterdam, The Netherlands, 2007; pp. 1–36. [CrossRef]
16. Sivaprakasam, P.; Angamuthu, M. Generalized Z-Fuzzy Soft B-Covering Based Rough Matrices and Its Application To Magdm Problem Based On Ahp Method. *Decis. Mak. Appl. Manag. Eng.* **2023**, *6*, 134–152. [CrossRef]
17. Ibrahim, H.; Anwar, S.A.; Ahmad, M.I. Classification of imbalanced data using support vector machine and rough set theory: A review. *J. Phys. Conf. Ser.* **2021**, *1878*, 12054. [CrossRef]
18. Juneja, M.; Walia, E.; Sandhu, P.S.; Mohana, R. Implementation and comparative analysis of rough set, Artificial Neural Network (ANN) and Fuzzy-Rough classifiers for satellite image classification. In Proceedings of the 2009 International Conference on Intelligent Agent & Multi-Agent Systems, Chennai, India, 22–24 July 2009; pp. 1–6. [CrossRef]
19. Albu, A.; Precup, R.E.; Teban, T.A. Results and challenges of artificial neural networks used for decision-making and control in medical applications. *Facta Univ. Ser. Mech. Eng.* **2019**, *17*, 285–308. [CrossRef]
20. Zhang, X.; Tian, Y.; Chen, L.; Hu, X.; Zhou, Z. Machine Learning: A New Paradigm in Computational Electrocatalysis. *J. Phys. Chem. Lett.* **2022**, *13*, 7920–7930. [CrossRef]
21. Omidvar, N.; Pillai, H.S.; Wang, S.H.; Mou, T.; Wang, S.; Athawale, A.; Achenie, L.E.; Xin, H. Interpretable Machine Learning of Chemical Bonding at Solid Surfaces. *J. Phys. Chem. Lett.* **2021**, *12*, 11476–11487. [CrossRef]
22. Pawlak, Z.; Skowron, A. Rudiments of rough sets. *Inf. Sci.* **2007**, *177*, 3–27. [CrossRef]
23. Pawlak, Z. Rough sets. *Int. J. Comput. Inf. Sci.* **1982**, *11*, 341–356. [CrossRef]
24. Mahajan, P.; Kandwal, R.; Vijay, R. Rough Set Approach in Machine Learning: A Review. *Int. J. Comput. Appl.* **2012**, *56*, 1–13. [CrossRef]
25. Aviso, K.B.; Janairo, J.I.B.; Promentilla, M.A.B.; Tan, R.R. Prediction of CO_2 storage site integrity with rough set-based machine learning. *Clean Technol. Environ. Policy* **2019**, *21*, 1655–1664. [CrossRef]
26. Chong, J.W.; Thangalazhy-Gopakumar, S.; Tan, R.R.; Aviso, K.B.; Chemmangattuvalappil, N.G. Estimation of fast pyrolysis bio-oil properties from feedstock characteristics using rough-set-based machine learning. *Int. J. Energy Res.* **2022**, *46*, 19159–19176. [CrossRef]
27. Heng, Y.P.; Lee, H.Y.; Chong, J.W.; Tan, R.R.; Aviso, K.B.; Chemmangattuvalappil, N.G. Incorporating Machine Learning in Computer-Aided Molecular Design for Fragrance Molecules. *Processes* **2022**, *10*, 1767. [CrossRef]
28. Cheun, J.-Y.; Liew, J.-Y.-L.; Tan, Q.-Y.; Chong, J.-W.; Ooi, J.; Chemmangattuvalappil, N.G. Design of Polymeric Membranes for Air Separation by Combining Machine Learning Tools with Computer Aided Molecular Design. *Processes* **2023**, *11*, 2004. [CrossRef]
29. Balaban, A.T. Highly discriminating distance-based topological index. *Chem. Phys. Lett.* **1982**, *89*, 399–404. [CrossRef]
30. Balaban, A.T.; Khadikar, P.V.; Supuran, C.T.; Thakur, A.; Thakur, M. Study on supramolecular complexing ability vis-à-vis estimation of pKa of substituted sulfonamides: Dominating role of Balaban index (J). *Bioorg. Med. Chem. Lett.* **2005**, *15*, 3966–3973. [CrossRef]
31. Pawlak, Z. Rough set theory and its applications. *J. Telecommun. Inf. Technol.* **2002**, *3*, 7–10. [CrossRef]
32. Vashist, R.; Vaishno, S.M.; Garg, M.L. Rule Generation based on Reduct and Core: A Rough Set Approach. *Int. J. Comput. Appl.* **2011**, *29*, 975–8887. [CrossRef]
33. Predki, B.; Słowiński, R.; Stefanowski, J.; Susmaga, R.; Wilk, S. ROSE—Software Implementation of the Rough Set Theory. In *Rough Sets and Current Trends in Computing*; Polkowski, L., Skowron, A., Eds.; Springer: Berlin/Heidelberg, Germany, 1998; pp. 605–608.

34. Grzymala-Busse, J.W. An Empirical Comparison of Rule Induction Using Feature Selection with the LEM2 Algorithm. In *Advances on Computational Intelligence*; Greco, S., Bouchon-Meunier, B., Coletti, G., Fedrizzi, M., Matarazzo, B., Yager, R.R., Eds.; Springer: Berlin/Heidelberg, Germany, 2012; pp. 270–279.
35. Balaban, A.T. Applications of Graph Theory in Chemistry. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 334–343. [CrossRef]
36. Bruckner, J.V.; Anand, S.S.; Warren, D.A. Toxic Effects of Solvents and Vapors. In *Casarett & Doull's Essentials of Toxicology*, 3rd ed.; Klaassen, C.D., Watkins, J.B., III, Eds.; McGraw-Hill Education: New York, NY, USA, 2015.
37. Kanu, I.; Anyanwu, E. Impact of hydrophobic pollutants' behavior on occupational and environmental health. *Sci. J.* **2005**, *5*, 211–220. [CrossRef] [PubMed]
38. Nikolić, S.; Trinajstić, N.; Mihalić, Z. The Wiener Index: Development and Applications. *Croat. Chem. Acta Ccacaa* **1995**, *68*, 105–129.