Explainable based approach for the air quality classification on the granular computing rule extraction technique

Idriss Jairi, Sarah Ben-Othman, Ludivine Canivet, Hayfa Zgaya-Biau

Explainable based Approach for the Air Quality Classification on the Granular Computing Rule Extraction Technique

Idriss Jairi[a] (idriss.jairi@univ-lille.fr), Sarah Ben-Othman[b]
(sara.ben-othman@centralelille.fr), Ludivine Canivet[c]
(ludivine.canivet@univ-lille.fr), Hayfa Zgaya-Biau[a]
(hayfa.zgaya-biau@univ-lille.fr)

[a] Univ. Lille, UMR 9189 - CRIStAL - Centre de Recherche en Informatique Signal et Automatique de Lille, F-59000 Lille, France

[b] Centrale Lille, UMR 9189 - CRIStAL - Centre de Recherche en Informatique Signal et Automatique de Lille, F-59000 Lille, France

[c] Univ. Lille, ULR 4515 - LGCgE, Laboratoire de Génie Civil et géo-Environnement, F-59000 Lille, France

Corresponding Author:
Idriss Jairi
Univ. Lille, UMR 9189 - CRIStAL - Centre de Recherche en Informatique Signal et Automatique de Lille, F-59000 Lille, France
Tel: +33 7 78 53 25 28
Email: idriss.jairi@univ-lille.fr

Explainable based Approach for the Air Quality Classification on the Granular Computing Rule Extraction Technique

Idriss Jairi[a,*], Sarah Ben-Othman[b], Ludivine Canivet[c], Hayfa Zgaya-Biau[a]

[a] *Univ. Lille, UMR 9189 - CRIStAL - Centre de Recherche en Informatique Signal et Automatique de Lille, F-59000 Lille, France*
[b] *Centrale Lille, UMR 9189 - CRIStAL - Centre de Recherche en Informatique Signal et Automatique de Lille, F-59000 Lille, France*
[c] *Univ. Lille, ULR 4515 - LGCgE, Laboratoire de Génie Civil et géo-Environnement, F-59000 Lille, France*

Abstract

Air pollution corresponds to one of the considerable challenges and disastrous sides of the environment that causes severe damage to all its biodiversity, including humans. As a result, establishing efficient, reliable, and interpretable methods and techniques to predict and control air quality is a must to preserve the environment and consider the necessary precautions. Most traditional machine learning models often lack transparency, making it challenging to interpret their decisions, especially in vital domains like air pollution. This paper proposes a novel approach that leverages granular computing to extract interpretable rules for air quality classification. We demonstrate the effectiveness of our approach through experiments on a real-world air quality dataset, showcasing the interpretability of the extracted rules and their accuracy in classifying air quality levels. The output of the proposed GrC model is a tree-like structure minimizing the entropy, allowing an easier interpretation of the classification results. A comparison is conducted with some widely used machine learning algorithms, including decision tree classifier, random forest, and CatBoost. The results in-

*Corresponding author
Email addresses: idriss.jairi@univ-lille.fr (Idriss Jairi), sara.ben-othman@centralelille.fr (Sarah Ben-Othman), ludivine.canivet@univ-lille.fr (Ludivine Canivet), hayfa.zgaya-biau@univ-lille.fr (Hayfa Zgaya-Biau)

dicate that the proposed granular computing rule extraction approach shows a competitive performance according to traditional black-box models in terms of accuracy (79%), transparency and reliability. The developed GrC model and the findings of this study not only contribute to advancing the field of air quality classification but also bear broader implications for environmental research and management for relevant and informed decision-making.

Keywords: Air quality classification, Granular computing, Rule extraction, Environmental monitoring, Machine learning

---

1. Introduction

Air pollution is one of the challenging issues and one of the major problems facing the environment and thus public health. It is due to the chemicals and particles that are suspended in the air from different sources, which pose serious threats to human health, ecosystems, and the overall quality of life. Some air pollution comes from natural such as wildfires, volcanic eruptions, or allergens. However, most air pollution results from human activities, including industrialization and the rapid development of urbanization. There are different types of air pollutants, including particulate matter ($PM_{2.5}$ and PM_{10}), sulfur dioxide (SO_2), nitrogen dioxide (NO_2), carbon monoxide (CO), carbon dioxide (CO_2), and ozone (O_3), and the concentrations of these air pollutants are usually measured using ambient air quality monitoring stations.

The impacts and consequences of air pollution are far-reaching. Exposure to polluted air (including different air pollutants) can lead to respiratory problems, cardiovascular diseases, lung diseases, cancer, and other diseases and illnesses (Kampa & Castanas, 2008). In addition, children, older people, and individuals with pre-existing health conditions are sensitive and particularly susceptible to the adverse effects of air pollution on their health. According to the WHO, ambient (outdoor) air pollution is estimated to have caused 4.2 million premature deaths worldwide in 2019 (WHO, 2022).

Realizing the severity of air pollution and its impact on human health and

environmental health in general, initiatives and efforts have been implemented to monitor, control, and reduce air pollution. Tackling the air pollution issue requires efforts from different sectors and approaches involving collective action, policy interventions, and technological advancements.

On the technological advancements side, and due to the increased volumes of collected data from the air quality monitoring stations, the more and more advanced algorithms, and the increase in computational power and storage, artificial intelligence (AI) and especially machine learning (ML) techniques, have become the tool of choice to monitor, forecast, and classify air quality for effective, relevant decision support (Noori et al., 2010; Moazami et al., 2016).

In this paper, we focus on the data-intensive machine learning technique: supervised learning. We distinguish two types of supervised learning applied for air quality; regression/forecasting and classification.

For forecasting (or regression), there are various machine learning algorithms and techniques that have been used to forecast either the air quality index (AQI) or the concentration of one of the air pollutants. Some of the widely employed techniques include autoregressive integrated moving average (ARIMA) (Kumar & Jain, 2010), recurrent neural network (RNN), LSTM (Tsai et al., 2018), and XGBoost (eXtreme Gradient Boosting) (Ma et al., 2020) (Section 2. Related works)

On the other hand, classifying the air quality, and especially extracting classification rules that explain the change in the levels of the air quality, have not been much explored. Classifying air quality and extracting classification rules to explain its changes are pivotal for multiple reasons. Primarily, they directly impact public health and environmental well-being. Extracting classification rules contributes to scientific knowledge by identifying complex relationships between meteorological variables, pollutants, and air quality levels and enhances our understanding of environmental processes.

In this paper, a granular computing (GrC) based method is proposed to extract air quality classification rules. The goal is to get intuitions about the factors that affect the changes in air quality. Similar studies conducted in the

other fields of science and engineering based on the GrC approach, such as estimation of the dispersion coefficient in natural rivers (Noori et al., 2017a), prediction of scour below spillways (Noori et al., 2017b), and prediction of pollutant longitudinal dispersion coefficient in aquatic streams (Ghiasi et al., 2022).

Granular computing (GrC) is a problem-solving and information paradigm that mimics human thinking and reasoning by dealing with information and knowledge in the form of aggregates called information granules. In general, granular computing refers to a comprehensive concept that covers theories, methodologies, techniques, and tools that make use of granules and perform computation on granules in complex problem-solving. In the context of granular computing, two key concepts are granules and granularity, the first one, granules, refers to a coherent and meaningful unit that encapsulates information, whereas, granularity denotes the level or scale at which information is organized, represented, or processed.

Granular computing is employed in our study due to its inherent capacity to handle complex and imprecise information through the formation of granules or clusters, which aids in capturing patterns and relationships within the data. Granules serve as a mechanism to simplify the representation of information, making it more manageable and interpretable. In the context of air quality classification, where datasets often exhibit intricate and complex patterns and dependencies, the granular approach allows us to represent data at different levels of abstraction. This hierarchical representation enables a more nuanced understanding of the relationships between meteorological features and air quality levels. Furthermore, the flexibility of granular computing aligns well with the nature of air quality data, which may involve uncertainties and variations. By leveraging granular computing, we aim to enhance the adaptability of our model to diverse and dynamic environmental conditions.

One of the applications of granular computing is classification rule extraction. The classification rule induction method focuses on selecting a single granule, instead of concentrating on selecting the suitable partition, which leads to finding a covering solution of the universe (Yao & Yao, 2002).

In this study, a heuristic algorithm is proposed to extract useful and meaningful classification rules that affect air quality levels using the granular computing approach. The proposed method aims to reduce the complexity of the data, making it easier to understand and work with. Moreover, this approach can provide a transparent and interpretable method for rule extraction, where the algorithm shows the reasoning behind the given model's decision-making. For example, high wind speed will reduce the concentration of $PM_{2.5}$, high humidity usually aggravates air pollution, and high atmospheric pressure usually results in good air quality.

The primary contribution of this study lies in the deployment of the granular computing (GrC) rule extraction approach as a novel methodology for air quality classification. The originality of the proposed approach is twofold. Firstly, the adoption of granular computing in the field of air quality classification introduces a promising paradigm shift from conventional machine learning models. In fact, Granular computing enables the extraction of rules in the form of granules, offering a more human-understandable representation of decision-making processes. This interpretability is crucial in environmental science. Secondly, the adaptability of the granular computing rule extraction algorithm, by proposing a new approach for rules extraction through minimizing the entropy of the granules.

The paper is structured as follows. A detailed literature review of the employed techniques in air quality forecasting and classification is illustrated and discussed in Section 2. This state-of-the-art allows the positioning of our work in relation to the existing literature and supports the choice of granular computing technique based on its demonstrated effectiveness in similar studies solving similar problems in various contexts and its alignment with the research objectives. Section 3 provides a definition and basic concepts of granular computing. Section 4 presents the proposed methodology and the steps involved in building the granular computing rule extraction classifier. The experimental results and comparisons to other machine learning models are detailed in Section 5. Section 6 presents a discussion of the findings, results interpretation, and limitations and

potential challenges of the proposed method. Finally, the last section (Section 7) summarizes the findings, results, improvements of the method, and future prospects.

2. Related works

There are several models and methods used to forecast and monitor the air quality. The deterministic models including Gaussian dispersion models are among the widely employed models to estimate the dispersion and transport of pollutants in the atmosphere (Abdel-Rahman, 2008). These models are based on the assumption that the dispersion of pollutants follows a Gaussian distribution, which means that the pollutant concentration decreases with distance from the source in a bell-shaped curve. Atmospheric Dispersion Modeling System (ADMS) (McHugh et al., 1997) is a widely-used air quality modeling system in Europe that is based on the Gaussian dispersion model.

3D Eulerian chemistry-transport models are another well-known type of deterministic model, that combines Eulerian (The Eulerian method treats the particle phase as a continuum and develops its conservation equations on a control volume basis and in a similar form as that for the fluid phase. (Zhang & Chen, 2007)) methods with detailed representations of atmospheric chemistry and transport processes. These models simulate the three-dimensional distribution of pollutants in the air and their interactions with meteorological conditions, emissions, and chemical reactions. For example, we can mention: CHIMERE (CHIMie-transport model for Emission and REgional scales) (Bessagnet et al., 2004; Menut et al., 2013) and CMAQ (Community Multiscale Air Quality) (Binkowski & Roselle, 2003). However, air quality is influenced by various factors and uncertainties, and the above-mentioned deterministic models have limitations and drawbacks in capturing all the complexities of atmospheric processes.

On the other hand, statistical, machine learning, and deep learning techniques are increasingly being used for air quality forecasting and monitoring due

to their ability to handle complex relationships and capture non-linear patterns in the data. Statistical regression models, such as linear regression, multiple linear regression, and generalized linear models, are used to establish relationships between air quality pollutant concentrations and relevant predictors, such as meteorological parameters, emission data, and historical pollutant levels (Slini et al., 2002; Kumar & Goyal, 2011). Time series analysis techniques, including autoregressive integrated moving average (ARIMA) models (Kumar & Jain, 2010; Abhilash et al., 2018) and seasonal decomposition of time series (STL), are used to capture temporal patterns and seasonality in air quality data. These models can identify trends, periodic fluctuations, and other time-dependent patterns in pollutant concentrations, aiding in short-term and long-term air quality forecasting. Other machine learning algorithms were used in air quality forecasting, such as Support Vector Machines (SVM), (Moazami et al., 2016) developed a support vector regression (SVR) model for predicting carbon monoxide concentration levels. Decision trees, random forests, and eXtreme Gradient Boosting (XGBoost) are among the widely used machine learning algorithms for air quality forecasting (Osowski & Garanty, 2007; Bozdağ et al., 2020; Castelli et al., 2020; Ma et al., 2020; Lei et al., 2023). Recently, researchers have started exploring deep-learning models including Artificial Neural Networks (ANNs) (Chelani et al., 2002; Niska et al., 2004; Kumar & Goyal, 2013), for example, (Noori et al., 2010) proposed a deep learning model to predict daily carbon monoxide(CO) concentration in the atmosphere of Tehran using artificial neural network (ANN) and adaptive neuro-fuzzy inference system(ANFIS).

Recurrent Neural Networks (RNNs) (Biancofiore et al., 2017; Zaini et al., 2022; Eren et al., 2023), and other sophisticated deep learning models; (Wang & Song, 2018) proposed a deep spatial-temporal ensemble(STE) model and LSTM, (Liu et al., 2019) forecasted four pollutants concentrations ($PM_{2.5}$, SO_2, NO_2, and CO) in Beijing, China, based on an intelligent hybrid model, and finally, (Du et al., 2019) proposed a deep learning model based on 1D-CNN and Bi-directional LSTM for $PM_{2.5}$ forecasting.

We focus in this paper on the development of a classifier model for classi-

fying air quality levels. Some classification models in the literature have been developed. Examples of the developed classifiers for air quality include (Zhao et al., 2013; Gore & Deshpande, 2017; Aggarwal et al., 2017; Teologo et al., 2018; Mangayarkarasi et al., 2021; Haq, 2022; Saminathan & Malathy, 2023). Table 1 is a literature summary table that provides a synopsis of the different reviewed articles for classifying levels of air quality.

Paper's reference	Method	Goal	Dataset	Shortcomings
(Kujaroentavon et al., 2014)	Decision tree	This research aims to establish rules of separated air quality classification (Classifying AQI)	The data were collected from the air quality pollution control department in Thailand in 2012-2013	Not considering meteorological data + The contribution was not stated + The results were not clearly explained
(Sugiarto & Sustika, 2016)	Decision Tree (C4.5 algorithm)	A classification algorithm is proposed for classifying air quality using the C4.5 algorithm. The entropy and information gain values are computed in order to construct the decision tree structure and build the rule sets	Experimental datasets were collected from sensor nodes	Very little training data was used + Considering few inputs
(Corani & Scanagatta, 2016)	A multi-label classifier, which simultaneously predicts multiple air pollution variables	A multi-label classifier based on Bayesian networks to predict $PM_{2.5}$ and Ozone	Shanghai data set which covers the period between February 2013 to February 2014 of 10 stations + The ozone data of Berlin, of one station in 1997-1999 + Burgas dataset for ozone the dataset eventually contains 208 daily recordings	None
(Gore & Deshpande, 2017)	Naive Bayes and Decision tree J48 Algorithms	The objective is to classify AQI categories based on the AQI of four pollutants	U.S. Pollution Data	Not considering meteorological data + The contribution was not stated + The quality of the paper and the developed models is below standards
(Aggarwal et al., 2017)	Fuzzy logic and fuzzy interface system	A fuzzy interface system for the calculation of AQI using two pollutants ($PM_{2.5}$ and PM_{10}) with each having six linguistic variables	Data for 5 days was collected from an open source	Considering only two pollutants ($PM_{2.5}$ and PM_{10})
(Teologo et al., 2018)	Fuzzy logic and Mamdani fuzzy inference system	A classification algorithm for the air quality index (AQI) using fuzzy logic (FL) system considering two pollutants (CO and NO_2)	Data were collected from an air quality monitoring portal (Philippines)	Considering only two pollutants (CO and NO_2)
(Zhao et al., 2018)	SVM, Random forest, and RNN	Predicting Daily Air Quality Classification in three cities in the US based on RNN model	Data is collected based on U.S. EPA for the period from January 1, 2010, to December 31, 2015, including a total of 2,191 observations	Not considering meteorological data + The results are not interpretable
(Hamami & Fithriyah, 2020)	Artificial neural network	This research proposes neural network methods to classify data into three air pollution levels	The IoT dataset is obtained from Open Data Jakarta, it contains 10 attributes with 1827 rows	Not considering meteorological data + The contribution was not stated + The results were not clearly explained
(Mangayarkarasi et al., 2021)	Logistic Regression and Random Forest	Classifying AQI Categories	World Air Quality Index historical data	Not considering meteorological data + The contribution is not clearly states + The results were not clearly explained
(Haq, 2022)	SMOTEDNN, XGBoost, Random Forest, SVM, and k-NN	A novel model SMOTEDNN to classify air pollution was developed and compared to four ML models, XGBoost, Random Forest, SVM, and k-NN.	The dataset was released under the NAMP program from Jan 01, 2015, to July 07, 2020.	Not considering meteorological data + The contribution was not clearly stated + the results seem to be too perfect
(Hamami & Dahlan, 2022)	Logistic Regression, KNN, Decision Tree, and Random Forest	Classifying air quality levels into three categories	The dataset was taken from Jakarta's open data for 12 months with several attributes	Not considering meteorological data + The contribution was not stated + The results were not clearly explained
(Saminathan & Malathy, 2023)	Logistic Regression, SVM, Random Forest, XGBoost, and Multi-layer perceptron	Classifying $PM_{2.5}$ values to different categories/groups	UCI Machine Learning Repository 2017	Application of existing approaches + The methodology is not clearly presented + The results are not clearly interpreted

Table 1: Air quality classification : literature overview

This review serves as a foundation for understanding the current state of

knowledge in the field of air quality classification and classification rules extraction and highlights the need for innovative and interpretable methodologies in both machine learning and environmental sciences fields. As mentioned in Table 1, there are not enough studies about classifying air quality and extracting interpretable rules that affect the levels of air pollutants and thus the air quality. Most of the studies have not included the meteorological features in the developed models, knowing that meteorological condition plays an important part in the levels of air quality (Jhun et al., 2015). Moreover, most of the studies have employed complex models that are not easy to understand and seem to be black boxes lacking interpretability and transparency. In contrast, granular computing focuses on generating coherent and understandable units of information, making it well-suited for the intricacies of air quality assessment.

The adoption of Granular Computing (GrC) has been applied by some research studies in different disciplines. (Ghiasi et al., 2022) developed an artificial intelligence-based predictive model, coupling granular computing and neural network models (GrC-ANN) and its uncertainty to provide robust estimation of pollutant longitudinal dispersion coefficient in aquatic streams. (Noori et al., 2017b) presented a new method for the prediction of the depth, length, and width of the scour hole downstream ski-jump buckets based on the granular computing (GrC) technique. (Noori et al., 2017a) explored a granular computing (GC) model for the first time to overcome problems of accurately estimating the dispersion coefficient in natural rivers. The innovative aspect of this approach lies in its ability to propose a new and effective algorithm that encapsulates domain-specific knowledge about meteorological features, and air pollutants, and extracts rules that are both accurate and comprehensible. Other conducted research works regarding the GrC deployment for rule extraction are presented in next section.

3. Granular computing

Granular computing (GrC) is a paradigm in information processing that aims to break down a complex problem into a bunch of sub-problems based on the similarity between granules or clusters (Bargiela & Pedrycz, 2006; Pedrycz, 2018; Bargiela & Pedrycz, 2022). The concept of granular computing was first called information granularity/granulation (Zadeh, 1979), while the term "Granular Computing" appeared for the first time in (Zadeh, 1997).

In granular computing, a system is decomposed into smaller entities or components called granules. These granules can be physical entities, conceptual objects, data points, or any other meaningful units that capture the essential features of the system. The granularity of a system refers to the size or scale of the granules used to represent it. The key idea behind granular computing is that different levels of granularity provide different perspectives and insights/knowledge into the system under study. By examining a system at multiple levels, researchers can capture both the macroscopic behavior and the microscopic details of the system, leading to a more comprehensive understanding.

Definition 1: Granular computing can be defined as an umbrella term to cover all theories, methodologies, techniques, and tools that make use of granules in complex problem-solving. The process of performing computation and operations on granules.

Definition 2: GrC is a general computation theory that imitates human thinking and reasoning by dealing with information as a form of aggregates called information granules.

Figure 1 illustrates the different levels of granularity/details in a scientific research paper. One can easily observe the different and multiple levels of processing information (granularity) in any scientific/technical writing.

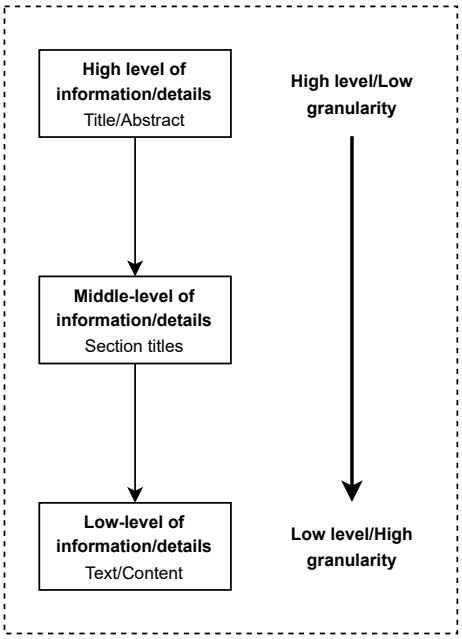


Figure 1: A simple illustration of different levels of granularity in a scientific paper

In GrC, two important notions are distinguished: granule and granularity. The term granule refers to a coherent and meaningful unit that encapsulates information. It represents a cluster of data that exhibits internal homogeneity and external heterogeneity. Granules can be thought of as building blocks that help organize and structure complex information. (The need for information abstraction). The term granularity refers to the level or scale at which information is organized, represented, or processed. It pertains to the degree of detail or coarseness in the division of data into granules. Granularity determines the size or extent of the individual granules and how they relate to each other in terms of their hierarchical or overlapping nature. (Knowledge of abstraction level).

Fuzzy sets (Zadeh, 1965) and rough sets (Pawlak, 1982) form the sound basis and the main driving forces of granular computing.

Granular computing can be applied in various domains including data mining, machine learning, pattern recognition, decision-making, and knowledge dis-

covery by dividing large datasets into meaningful granules, it becomes easier to identify relevant patterns, extract main features, and discover hidden knowledge. The GrC can be used in decision support systems by processing data/granules at different levels of granularity, in order to handle uncertainty, ambiguity, and incomplete information.

One of its key applications is in the extraction of classification rules. This involves deriving rules that can be used to classify or categorize data into different classes or categories based on specific attributes or features. This process is widely used in fields such as machine learning, data mining, and pattern recognition to make predictions or decisions based on available data.

The next subsections illustrate the foundation of constructing a granular computing model for classification rules extraction.

3.1. Information table

An information table is a fundamental concept used for representing and organizing data. It is a tabular structure that contains information about objects or entities, their attributes, and their corresponding attribute values. The information table can be represented as follows:

$$S = (U, A_t, L, \{V_a | a \in A_t\}, \{f_a | a \in A_t\}) \tag{1}$$

where U is a non-empty set of objects, A_t is a non-empty set of attributes, V_a is a set of values for the attribute a, where $a \in A_t$, and f_a is an information function that maps each element from the universe U to a value V_a, f_a: U $\rightarrow$ V_a. L is the defined language for the attributes A_t, where an atomic formula is given by a = v, a $\in A_t$, and v $\in V_a$.

Granulation of the universe U is the process of dividing the objects of U into clusters, groups, or subsets based on the similarity between these objects, where each of these subsets or groups is called a granule. A granule may be viewed as a subset of the universe, which maybe either fuzzy or crisp (Yao, 2004). Granules can be created and constructed from the language L. For an atomic formula a = v, we obtain the basic granule m(a = v) (Yao & Yao, 2002).

3.2. Granules measures and evaluation metrics

To extract classification rules, several evaluation metrics are used to evaluate and measure the granule as well as the relationship between a pair of granules. The first measure is Generality (Equation 2), which measures a single granule of formula (ϕ) by dividing the total number of elements/objects of the granule $m(\phi)$, by the total number of elements in the universe U.

$$Generality(\phi) = \frac{|m(\phi)|}{|U|} \tag{2}$$

The second measure is called Absolute Support or Confidence (Equation 3), which measures and quantifies the strength of two formulas ϕ and ψ ($\phi \implies \psi$)

$$Confidence(\phi \rightarrow \psi) = \frac{|m(\phi \wedge \psi)|}{|m(\phi)|} = \frac{|m(\phi \cap \psi)|}{|m(\phi)|} \tag{3}$$

Another measure to quantify the strength of the two formulas ϕ and ψ is called Coverage (Equation 4).

$$Coverage(\phi \rightarrow \psi) = \frac{|m(\phi \wedge \psi)|}{|m(\psi)|} \tag{4}$$

The last measure and the most important one, Conditional Entropy (Equation 5), which measures the homogeneity of the objects in the granule, is defined as follows:

$$ConditionalEntropy(\psi|\phi) = -\sum_{i=1}^{n} P(\psi_i|\phi) \log(P(\psi_i|\phi)) \tag{5}$$

Considering the family of formulas ψ =$\{\psi_1,...,\psi_n\}$ and let $\phi \implies \psi$ represent the inference relation between the formulas ϕ and ψ. The probability distribution P of $\phi \implies \psi$ is computed using the Equation 5.

3.3. Induction of classification rules and granule tree construction

The first step includes creating a family of basic concepts based on the atomic formulas (a = v). Secondly, the basic granules are measured and evaluated

using the metrics mentioned in the Subsection 3.2. Based on these metrics and by minimizing the conditional entropy and maximizing the other metrics (generality, confidence, and coverage) the rules can be extracted. The Algorithm 1 below illustrates the basic steps to construct the granule network (Yao & Yao, 2002).

Algorithm 1 Granule Network Construction

Construct the family of basic concepts with respect to atomic formulas: $BC(U) = (a = v, m(a = v))|a \in A_t, v \in V_a$.

Set the unused basic concepts to the set of basic concepts: $UBC(U) = BC(U)$.

Set the granule network to $GN = (U, \emptyset)$, which is a graph consisting of only one node and no arcs.

while the set of smallest granules in GN is not a covering solution of the classification problem **do**

Compute the fitness of each unused basic concept.

Select the basic concept $C = (a = v, m(a = v))$ with maximum value of fitness (minimum entropy and maximum generality, confidence, and coverage).

Set $UBC(U) = UBC(U) - C$.

Modify the granule network GN by adding new nodes which are the intersection of $m(a = v)$ and the original nodes of GN; connect the new nodes by arcs labelled by $a = v$.

The induction of classification rules using granular computing (GrC) has been applied by a few researchers and research studies in various disciplines. (Rozehkhani & Mohammadzad, 2022) used granular computing to classify patients and diagnosing COVID-19 disease by symptoms. (Samadi Alinia & Delavar, 2011; Khamespanah et al., 2013; Sheikhian et al., 2017) employed the granular computing approach to classify and assess seismic vulnerability. (Rozehkhani & Mahan, 2022) applied GrC for rules extraction to compute the

number of virtual machines based on some related features.

In general, there has not been much work to develop and apply GrC for rule extraction. This study developed the GrC algorithm to extract meteorological rules affecting $PM_{2.5}$ levels from scratch. To the best of our knowledge, the developed algorithms in the literature are not well constructed and explained, and they do not work for all cases, moreover, the algorithms have been tested on small datasets.

4. Methodology

In this section, we present the proposed approach to derive and extract classification rules for air quality levels ($PM_{2.5}$ levels: Good, Poor, and Extremely Poor). Figure 2 presents the adopted methodology process.

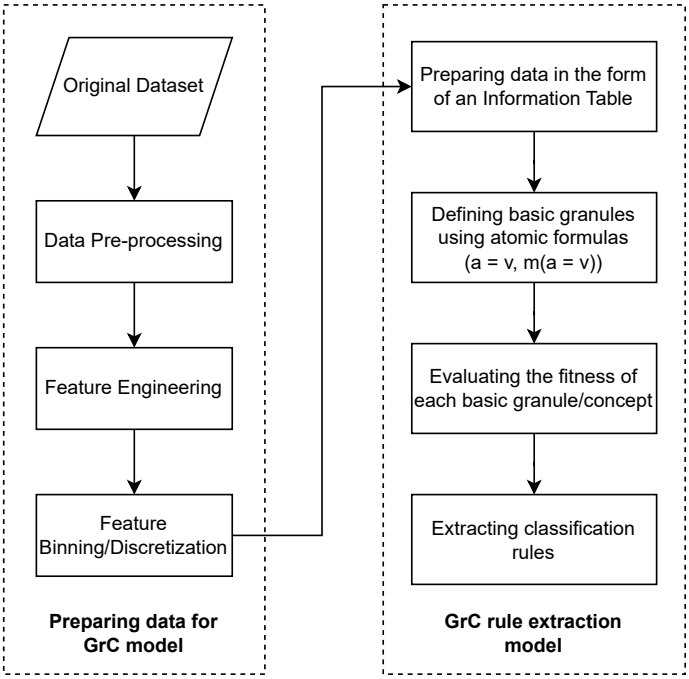


Figure 2: Overview of the proposed GrC model flowchart

As it is shown in Figure 2, the proposed model consists of two important

parts. The first involves preparing the data in a suitable format for the GrC model. The second part includes three significant steps, first, defining the basic granules in the form of 'attribute = value', second, evaluating and measuring the fitness of the specified basic granules/concept, and third, extracting classification rules, which is the most complex part, in which a heuristic algorithm is proposed in this study to extract rules with minimum entropy and maximum coverage values.

4.1. Data preparation and Information table production

The employed dataset is the Beijing $PM_{2.5}$ Data (Chen, 2017). It includes the hourly data of $PM_{2.5}$ concentrations and seven meteorological variables: temperature, pressure, dew point temperature, combined wind direction, cumulated wind speed, cumulated hours of snow, and cumulated hours of rain. It concerns the time period from January 1st, 2010 to December 31st, 2014, with a total of 43824 instances and 13 attributes/features. Table 2 summarizes the details of the attributes included in the dataset.

Column	Description	Non-Null Count	Dtype
No	row number	43824 non-null	int64
Year	year of data in this row	43824 non-null	int64
Month	month of data in this row	43824 non-null	int64
Day	day of data in this row	43824 non-null	int64
Hour	hour of data in this row	43824 non-null	int64
$PM_{2.5}$	$PM_{2.5}$ concentration (ug/m^3)	41757 non-null	float64
DEWP	Dew Point (°C)	43824 non-null	int64
TEMP	Temperature (°C)	43824 non-null	float64
PRES	Pressure (hPa)	43824 non-null	float64
CBWD	Combined wind direction	43824 non-null	str
Iws	Cumulated wind speed (m/s)	43824 non-null	float64
Is	Cumulated hours of snow	43824 non-null	int64
Ir	Cumulated hours of rain	43824 non-null	int64

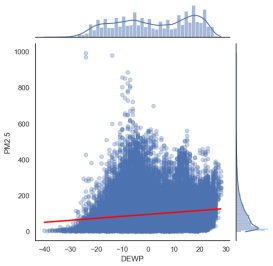
Table 2: A summary of the Beijing $PM_{2.5}$ Data attributes

The first step includes pre-processing the data. Data pre-processing involves a series of crucial steps aimed at enhancing the quality and suitability of raw data for subsequent analysis. This typically includes data cleaning to identify and rectify errors, missing values, and outliers. In the Beijing $PM_{2.5}$ data, the $PM_{2.5}$ column is the only column that contains null values (NaN) in almost 2067 instances (4.71% of the data), which were subsequently dropped.

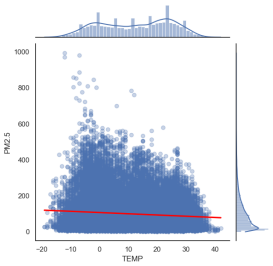
Feature selection or extraction is performed to identify relevant attributes and reduce dimensionality. The objective of this study is to determine how the levels of meteorological features affect the $PM_{2.5}$ levels. To achieve this, five meteorological features were chosen, namely, DEWP, TEMP, PRES, CBWD, and Iws as the main features/attributes to use for building the main GrC model and therefore extract the meteorological rules that affect the levels of $PM_{2.5}$. The reason behind dropping the two features Is and Ir is that most of the values

are set to 0 (99% for Is and 95% for Ir).

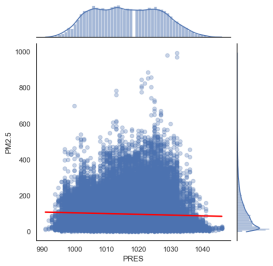
Figure 3 represents the bivariate analysis between the meteorological features and the levels of $PM_{2.5}$. Figures 3a, 3b, 3c, 3d, 3e, and 3f, examine the trend and association between the numerical meteorological variables (DEWP, TEMP, PRES, Iws, Ir, and Is) and the output column ($PM_{2.5}$) through a joint plot featuring a regression line. On the other hand, the two remaining plots (Figures 3g and 3h) represent the strip and box plots, illustrating the distribution and variation of $PM_{2.5}$ across different categories of combined wind directions (CBWD).



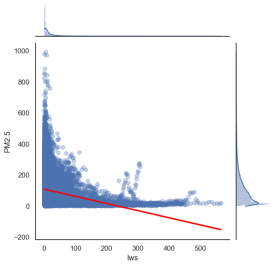
(a) Relationship between DEWP and $PM_{2.5}$



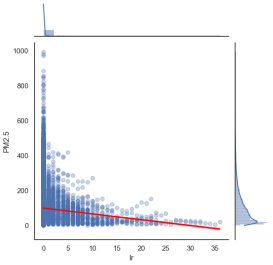
(b) Relationship between TEMP and $PM_{2.5}$



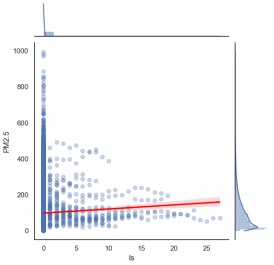
(c) Relationship between PRES and $PM_{2.5}$



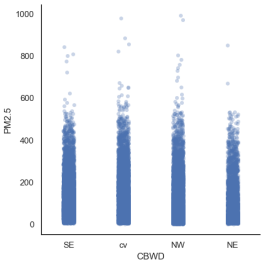
(d) Relationship between Iws and $PM_{2.5}$



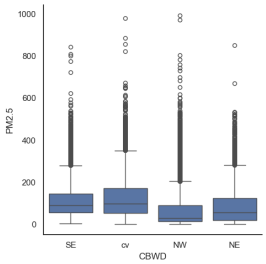
(e) Relationship between Ir and $PM_{2.5}$



(f) Relationship between Is and $PM_{2.5}$



(g) Relationship between CBWD and $PM_{2.5}$ (Strip plot)



(h) Relationship between CBWD and $PM_{2.5}$ (Box plot)

Figure 3: Bivariate analysis of the meteorological features and the output ($PM_{2.5}$)

As illustrated in Table 2, there are seven meteorological features versus the $PM_{2.5}$ concentration. The objective is to extract the rules that affect the different levels of $PM_{2.5}$. Figure 4 represents the correlation matrix of the columns in

the dataset, including the correlation between the meteorological features and the target ($PM_{2.5}$). It corresponds to the Pearson's correlation between two variables X and Y, Equation 6, where $\bar{X}$ and $\bar{Y}$ represent the mean values of X and Y respectively.

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2}} \tag{6}$$

As illustrated in the figure, there is a strong correlation between some meteorological features (e.g. DEWP and PRES, TEMP and PRES), which is normal. On the other hand, there are no strong correlations (linear correlations) between the output ($PM_{2.5}$) and meteorological features (DEWP, TEMP, PRES, Iws, Is, and Ir). However, the absence of a strong linear correlation does not imply that there is no relationship between the levels of $PM_{2.5}$ and the meteorological features. Therefore, these meteorological features are considered to extract meaningful rules about the meteorological variables that affect the levels of $PM_{2.5}$.

One crucial aspect to consider is the concentration level of $PM_{2.5}$ at time t is surely affected by the concentration level at t-n. For this, ACF (Auto-Correlation Function) and PACF (Partial AutoCorrelation Function) are two important notions for measuring how each data point in a series relates to its past data points. ACF measures how each data point in a series relates to its past data points including the indirect correlations in the calculation. It helps us understand if there's a pattern or trend that repeats at certain intervals. PACF, on the other hand, focuses on the direct relationship between two data points, while ignoring the influence of the other data points in between. It helps us to find out how one data point is directly connected to another, without the "influence" of the data points in the middle. Figure 5 shows the ACF and PACF plots of $PM_{2.5}$ for 40 lags/hours.

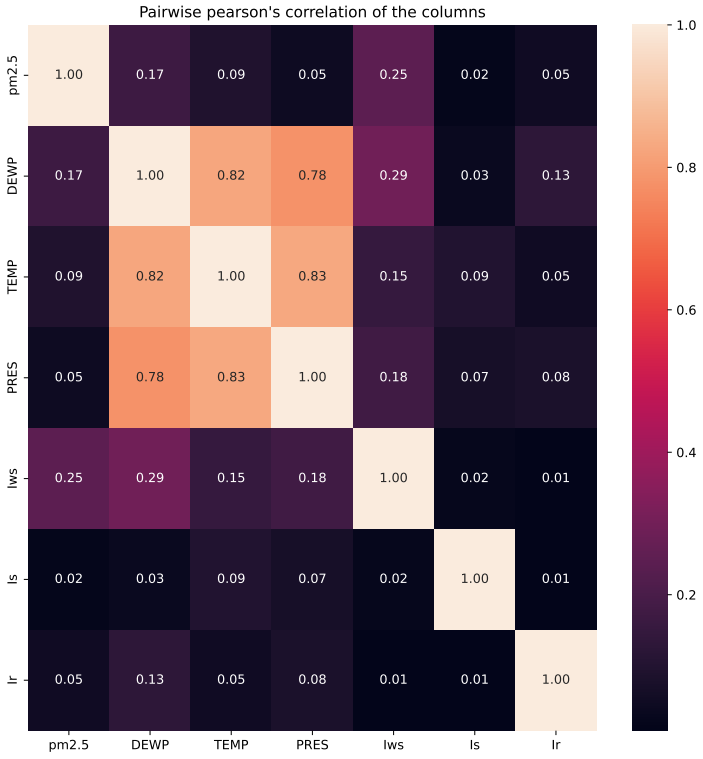


Figure 4: Pairwise Pearson's correlation of the columns

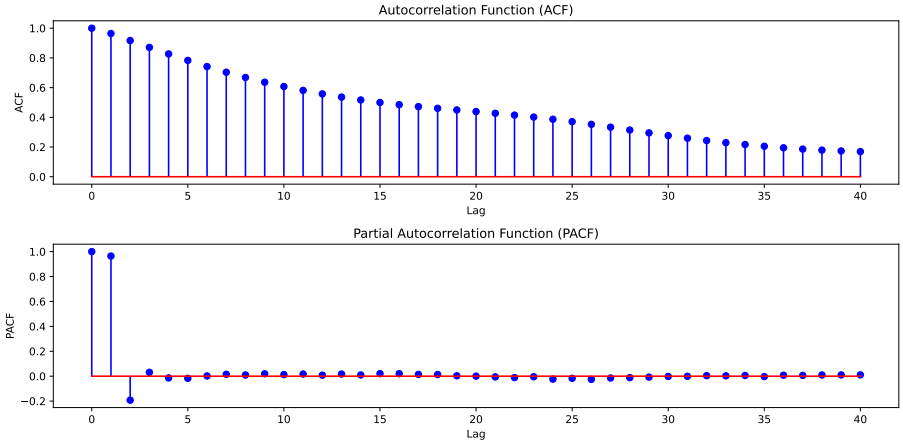


Figure 5: $PM_{2.5}$ ACF and PACF plots

Considering the PACF plot, there is a strong correlation between t and t-1, therefore, for this study, the one lag value for $PM_{2.5}$ is considered as an additional feature/attribute.

In this paper, our proposed GrC algorithm is based on feature binning (categorical columns), also known as feature discretization, which is a technique used in data preprocessing to transform continuous numerical features into discrete bins or intervals. This process simplifies the data and can make it easier to analyze and use with the GrC model. Discretization can simplify the model by converting continuous features into categorical ones, with less impact of outliers, and the possibility to capture non-linear relationships between features and target variables. Binning can be beneficial in different cases and situations where discretizing a feature for interpretability is desired, and that's the goal of this study, extracting and understanding the conditions that heavily affect the $PM_{2.5}$ level.

Table 3 shows the attributes (features and target) of the information table and the corresponding values/categories for each point.

Temperature	Pressure	Dew-point	Wind Speed	Wind Direction	One Lag $PM_{2.5}$	$PM_{2.5}$
Very cold = $]-\infty, 0]$	Very low = $]-\infty, 980]$	Very dry = $]-\infty, 0]$	Calm = $[0, 0.5]$	NE	Good = $[0, 50]$	Good = $[0, 50]$
Cold = $]0, 10]$	Low = $]980, 1000]$	Dry = $]0,9]$	Light air = $]0.5, 1.5]$	SE	Poor = $]50, 300]$	Poor = $]50, 300]$
Cool = $]10, 20]$	Normal = $]1000, 1020]$	Comfortable = $]9, 15]$	Gentle breeze = $]1.5, 5]$	NW	Extremely poor = $]300, +\infty[$	Extremely poor = $]300, +\infty[$
Warm = $]20, 30]$	High = $]1020, 1030]$	Slightly uncomfortable = $]15, 20]$	Fresh breeze = $]5, 10.5]$	CV		
Hot = $]30, +\infty[$	Very high = $]1030, +\infty[$	Moderately uncomfortable = $]20, 23]$	Strong breeze = $]10.5, 13.5]$			
		Extremely uncomfortable = $]23, +\infty[$	Moderate gale = $]13.5, 20]$			
			Strong gale = $]20, 27]$			
			Violent storm = $]27, +\infty[$			

Table 3: Attributes and corresponding values

A sample of the final dataset (Information table, Section 3) is illustrated in Table 4, including six features (One lag $PM_{2.5}$, TEMP, PRES, DEWP, Iws, and CBWD) and one output $PM_{2.5}$ level.

Row ID	One lag $PM_{2.5}$	TEMP	PRES	DEWP	Iws	CBWD	$PM_{2.5}$ Level
821	Good	Cool	Low	Slightly Uncomfortable	Gentle Breeze	CV	Good
1123	Good	Cool	Normal	Slightly Uncomfortable	Moderate Gale	NW	Good
213	Poor	Warm	Normal	Moderately Uncomfortable	Gentle Breeze	CV	Poor
1182	Poor	Hot	Normal	Comfortable	Light Air	CV	Poor
136	Poor	Very cold	High	Very Dry	Gentle Breeze	CV	Poor
773	Poor	Cold	Normal	Very Dry	Strong Breeze	CV	Poor
1071	Good	Cool	Normal	Very Dry	Light Air	NE	Good
755	Poor	Hot	Low	Very Dry	Violent Storm	NW	Good
477	Good	Warm	Low	Slightly Uncomfortable	Light Air	CV	Poor
387	Poor	Cool	High	Very Dry	Strong Breeze	NE	Good

Table 4: A sample from the final Beijing $PM_{2.5}$ dataset (Information table)

In conclusion, the data preprocessing phase forms a critical foundation for our analysis. Through rigorous data cleaning and preparation, we have ensured the accuracy and integrity of our dataset. Thoughtful feature selection has enabled us to focus on relevant attributes, enhancing the efficiency of subsequent analyses. The determination of an optimal lag length has been instrumental in capturing meaningful temporal patterns within the data. Additionally, the implementation of feature binning has facilitated the transformation of continuous variables into discrete categories, simplifying their interpretation and potential impact on the GrC model. By executing these preprocessing steps carefully, we have established a solid basis for extracting insightful patterns and building a robust GrC model.

4.2. The proposed AirQ-RuleGrCEx algorithm: Air Quality Rule Extraction through Granular Computing

Once the initial critical task of data preparation has been accomplished, the subsequent stage involves constructing a heuristic algorithm that extracts classification rules from the prepared data.

4.2.1. Basic granules table and atomic formulas

The first step is to build the basic granules table using the information table and atomic formulas under the form (attribute = value, m(attribute = value)), where m(attribute = value) represents the objects (granule) that satisfy the (attribute = value) rule. It is noteworthy that in this study the granulation process is done by using crisp sets as illustrated in Table 3, fuzzy sets can also be used as another alternative. The second step includes measuring these basic granules by computing the generality, confidence, coverage, and entropy (Equations 2, 3, 4, and 5). Table 5 shows the basic granules for the sample of data mentioned in Table 4, and Table 6 presents the measurements of each formula/granule.

Formula	Granule
One lag $PM_{2.5}$ = Good	[821, 1123, 1071, 477]
One lag $PM_{2.5}$ = Poor	[213, 1182, 136, 773, 755, 387]
TEMP = Cool	[821, 1123, 1071, 387]
TEMP = Warm	[213, 477]
TEMP = HOT	[1182, 755]
TEMP = Very cold	[136]
TEMP = Cold	[773]
PRES = Low	[821, 755, 477]
PRES = Normal	[1123, 213, 1182, 773, 1071]
PRES = High	[136, 387]
DEWP = Slightly uncomfortable	[821, 1123, 477]
DEWP = Moderately uncomfortable	[213]
DEWP = Comfortable	[1182]
DEWP = Very dry	[136, 773, 1071, 755, 387]
Iws = Gentle breeze	[821, 213, 136]
Iws = Moderate gale	[1123]
Iws = Light air	[1182, 1071, 477]
Iws = Strong breeze	[773, 387]
Iws = Violent storm	[755]
CBWD = CV	[821, 213, 1182, 136, 773, 477]
CBWD = NW	[1123, 755]
CBWD = NE	[1071, 387]

Table 5: Basic granules of data in Table 4

Formula	Granule	Generality	Confidence			Coverage			Entropy
			Class 0	Class 1	Class 2	Class 0	Class 1	Class 2	
One lag $PM_{2.5}$ = Good	[821, 1123, 1071, 477]	0.4	0.75	0.25	0.0	0.6	0.2	0.0	0.244
One lag $PM_{2.5}$ = Poor	[213, 1182, 136, 773, 755, 387]	0.6	0.333	0.666	0.0	0.4	0.8	0.0	0.27
TEMP = Cool	[821, 1123, 1071, 387]	0.4	1.0	0.0	0.0	0.8	0.0	0.0	0.0
TEMP = Warm	[213, 477]	0.2	0.0	1.0	0.0	0.0	0.4	0.0	0.0
TEMP = HOT	[1182, 755]	0.2	0.5	0.5	0.0	0.2	0.2	0.0	0.30
TEMP = Very cold	[136]	0.1	0.0	1.0	0.0	0.0	0.2	0.0	0.0
TEMP = Cold	[773]	0.1	0.0	1.0	0.0	0.0	0.2	0.0	0.0
PRES = Low	[821, 755, 477]	0.3	0.66	0.33	0.0	0.4	0.2	0.0	0.27
PRES = Normal	[1123, 213, 1182, 773, 1071]	0.5	0.4	0.60	0.0	0.4	0.6	0.0	0.29
PRES = High	[136, 387]	0.2	0.50	0.5	0.0	0.2	0.2	0.0	0.30
DEWP = Slightly uncomfortable	[821, 1123, 477]	0.3	0.66	0.33	0.0	0.4	0.2	0.0	0.27
DEWP = Moderately uncomfortable	[213]	0.1	0.00	1.00	0.0	0.0	0.2	0.0	0.0
DEWP = Comfortable	[1182]	0.1	0.0	1.0	0.0	0.0	0.2	0.0	0.00
DEWP = Very dry	[136, 773, 1071, 755, 387]	0.4	0.50	0.50	0.0	0.4	0.4	0.0	0.30
Iws = Gentle breeze	[821, 213, 136]	0.3	0.33	0.66	0.0	0.2	0.4	0.0	0.27
Iws = Moderate gale	[1123]	0.1	1.0	0.0	0.0	0.2	0.0	0.0	0.0
Iws = Light air	[1182, 1071, 477]	0.3	0.33	0.66	0.0	0.2	0.4	0.0	0.27
Iws = Strong breeze	[773, 387]	0.2	0.50	0.50	0.0	0.2	0.2	0.0	0.30
Iws = Violent storm	[755]	0.1	1.0	0.0	0.0	0.2	0.0	0.0	0.00
CBWD = CV	[821, 213, 1182, 136, 773, 477]	0.6	0.166	0.833	0.0	0.2	1.0	0.0	0.19
CBWD = NW	[1123, 755]	0.2	1.0	0.00	0.0	0.4	0.0	0.0	0.0
CBWD = NE	[1071, 387]	0.2	1.00	0.00	0.0	0.4	0.0	0.0	0.0

Table 6: Basic granules and their measures

The second part concerns developing a heuristic algorithm to extract classification rules. Two main sets are defined. The first one is the covering solution set (Equation 7) which is set to be empty in the beginning since there are no classified objects yet, and the second one in the remaining objects set (Equation 8), which contains the remaining objects that are not classified yet, initially, this set will contain all the objects in the data frame/information table.

$$covering\_solution = \{\emptyset\} \tag{7}$$

$$remaining\_objects = set(information\_table) - set(covering\_solution) \tag{8}$$

The conditional entropy (Equation 5) and generality (Equation 2) form the sound basis to select the optimal rules through minimizing entropy and maximizing generality. Entropy measures the impurity of a granule, which means that, the smaller the entropy, the lower the randomness/impurity. An entropy value of zero means that the objects of the granule belong to the same class, whereas, generality measures the ratio of the objects in the granule.

4.2.2. The proposed algorithm

The first step of the proposed algorithm includes setting a min-entropy threshold value. The threshold in this paper is set to zero, in order to extract all the rules that have the min-entropy value. Once the formulas/rules are extracted for the min-entropy threshold, we update the covering solution in Equation 7, by adding the extracted objects with the min-entropy threshold. The remaining objects set (Equation 8) is also updated. Once the min-entropy threshold formulas are extracted, there might be some remaining objects that are not classified during the first step which is extracting only the formulas (therefore objects) that satisfy the min-entropy threshold, and the question is: out of all the remaining basic formulas/granules how to choose the best formula taking into consideration two criteria. The first criterion aims to prevent duplicating the classification of objects already processed and classified. The second criterion focuses on finding the rule that contains the maximum number of objects in the remaining objects list. To solve this, we propose using the Jaccard Index (Equation: 9) which measures the similarity for the two sets of data A and B, with a range from 0% to 100%. The higher the percentage, the more similar the two sets are. In our case, 'A' represents the set of granules of the remaining formulas, and 'B' represents the set of the remaining objects.

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} \tag{9}$$

This recursive algorithm/process stops once all the remaining objects are processed and classified. The overall process of the proposed method for constructing a granular decision tree is illustrated in Algorithm 2. Moreover, for better explanation of the proposed algorithm, Figure 6 represents a step-by-step flowchart of the proposed GrC rule extraction algorithm.

Algorithm 2 The AirQ-RuleGrCEx algorithm

procedure GRANULENETBUILDING(inf_table)

Get the information table inf_table

Construct the family of basic concept with respect to atomic formulas (a = v, m(a = v))

Compute fitness (generality, confidence, coverage, and entropy) for the basic granules/formulas

Get the granules/formulas with min-entropy value (granules_min_entropy)

for *granule* in *granules_min_entropy* **do**

if granule['entropy_val'] $\neq$ 0 **then**

infTable $\leftarrow$ information_table[granule]

GranuleNetBuilding(infTable)

else

Update the covering_solution by adding the objects in the *granule*

Update the remaining_objs

remaining_objs $\leftarrow$ inf_table - covering_solution

while *remaining_objs* $\neq \{\emptyset\}$ **do**

Get the granule that has the highest Jaccard Index value with the remaining objects '*remaining_objs*' as granule

Update covering solution by adding the objects in the granule that has the highest Jaccard index with the remaining objects '*remaining_objs*'

Update the reaming objects '*remaining_objs*'

infTable $\leftarrow$ information_table[granule]

GranuleNetBuilding(infTable)

Applying the algorithm 2 to the demo data presented in the previous sections (Tables: 4, 5, 6), a granular tree is built as illustrated in Figure 7. As illustrated in Figure 7 the output model is a tree-like structure (interpretable flowchart) with a minimum entropy value.

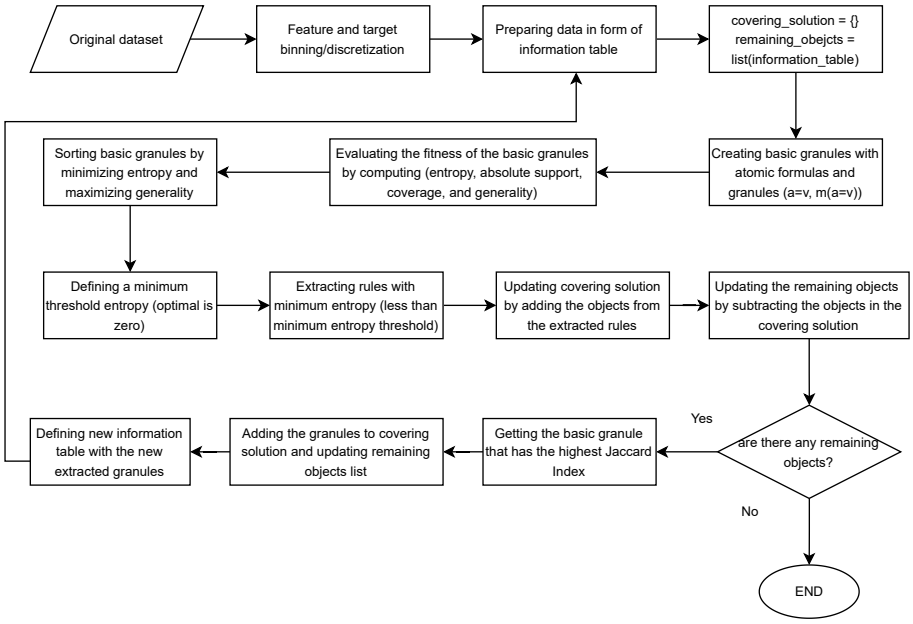


Figure 6: The proposed GrC rule extraction algorithm flowchart

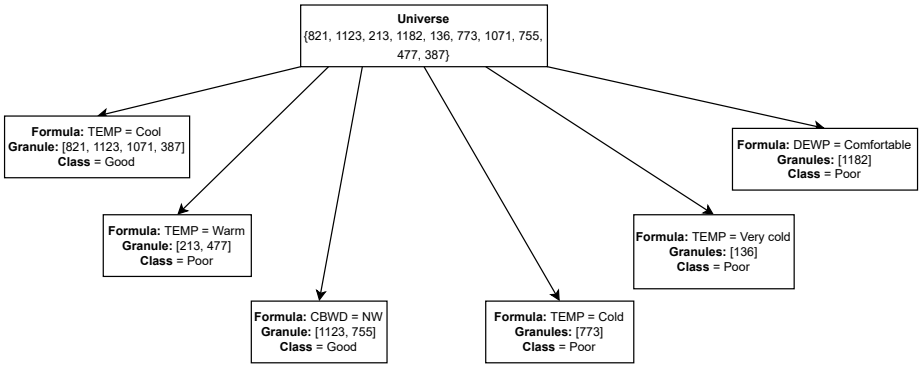


Figure 7: Granular tree demo

The methodology presented here offers a comprehensive approach for the preparation of data to facilitate its effective utilization within the Granular Computing (GrC) model and the steps involved in developing the GrC-base heuristic algorithm for air quality rules extraction. The initial phase involves the careful selection of pertinent features and their subsequent preprocessing,

including handling missing values and feature discretization. Subsequently, the GrC model starts the complex task of creating rules, systematically by creating basic granules under the form (attribute = value), evaluating the fitness of basic granules (generality, confidence, coverage, and entropy), and finally proposing and developing a sophisticated algorithm for air quality rule generation. This methodology offers a comprehensive solution for deriving valuable insights from complex datasets which can be improved in future studies.

5. Experiments and results

5.1. Practical applications

The practical applications of the proposed granular computing (GrC) methodology in air quality classification can be considered as an emerging machine learning model (tree-based model), that can be used besides the existing tree-based machine learning models (e.g. decision tree, random forest). By employing GrC for rule extraction, our method offers a robust framework for enhancing air quality management systems. The extracted rules provide a clear and interpretable foundation for understanding the relationships between meteorological conditions and air quality levels. This, in turn, empowers environmental monitoring initiatives with more accurate insights. Furthermore, the adaptability of the GrC algorithm allows it to be integrated seamlessly into existing air quality monitoring frameworks. This not only improves the accuracy of real-time air quality assessments but also provides a valuable tool for policymakers and public health officials in implementing targeted interventions. Overall, the proposed method stands as a promising advancement with direct implications for the efficacy of air quality management and, consequently, the well-being of communities.

5.2. Evaluation of the proposed model

In this section, we present a comprehensive set of experiments conducted to evaluate the effectiveness of the proposed granular computing approach for air

quality classification. The experiments are carried out on the Beijing $PM_{2.5}$ Data (Chen, 2017) dataset to validate the developed model and to extract the air quality classification rules that affect the levels of $PM_{2.5}$. The final dataset comprises 1801 rows representing $PM_{2.5}$ level categories (Good, Poor, and Extremely poor). The dataset is partitioned into training and testing sets using an 80-20 split. Out of 1441 training rows, a total of 588 rules have been extracted. Table 7 represents a sample of the extracted rules with the outcome for each rule.

Rule ID	Conditions	Outcome
264	PRES = Normal & $PM_{2.5}$_lagged_1 = Extremely poor & CBWD = NE & Iws = Gentle breeze & TEMP = Very cold	Poor
300	PRES = Low & $PM_{2.5}$_lagged_1 = Good & TEMP_level = Warm & CBWD = NW & DEWP = Dry & Iws = Moderate gale	Good
519	PRES = High & DEWP = Comfortable & TEMP = Warm & $PM_{2.5}$_lagged_1 = Good & CBWD = CV	Poor
212	PRES = Normal & $PM_{2.5}$_lagged_1 = Good & CBWD = CV & Iws = Gentle breeze & TEMP = Cold	Good
67	PRES_level = Normal & $PM_{2.5}$_lagged_1 = Poor & TEMP_level = Hot & CBWD = NW & DEWP = Comfortable & Iws = Moderate gale	Good
339	PRES = Low & $PM_{2.5}$_lagged_1 = Poor & TEMP = Warm & CBWD = SE & Iws = Moderate gale & Poor	
468	PRES = High & DEWP = Dry & $PM_{2.5}$_lagged_1 = Good & CBWD = NE & TEMP = Warm	Good
343	PRES = Low & $PM_{2.5}$_lagged_1 = Poor & TEMP = Warm & CBWD = SE & Iws = Violent storm	Poor
20	PRES = Normal & $PM_{2.5}$_lagged_1 = Poor & TEMP = Cool & DEWP =Dry & Iws = Fresh breeze	Poor
269	PRES_level=Normal & $PM_{2.5}$_lagged_1 = Extremely poor & CBWD = CV & TEMP = Warm & DEWP = Comfortable	Extremely poor

Table 7: Some extracted rules from the training set

Table 8 provides a comprehensive overview of the GrC classification model's performance on both the training and test sets. The metrics evaluated in-

clude accuracy, precision, recall, and f1-score which collectively offer insights into the model's ability to accurately classify air quality levels. Evidently, the outcomes for the training set exhibit 100% accuracy, which can be attributed to the model's successful extraction of all rules within this subset. Conversely, the test set demonstrates good performance, showcasing the model's ability to generalize its learned rules effectively.

Model	Accuracy	Class	Metrics		
			Precision	Recall	F1-score
Training set (80%)	1.00	Good (Class 0)	1.00	1.00	1.00
		Poor (Class 1)	1.00	1.00	1.00
		Extremely poor (Class 2)	1.00	1.00	1.00
Test set (20%)	0.79	Good (Class 0)	0.82	0.81	0.81
		Poor (Class 1)	0.76	0.80	0.78
		Extremely poor (Class 2)	0.85	0.71	0.77

Table 8: Results of the GrC model on training and test sets

To assess the robustness and the performance of the developed model on unseen data, the k-fold cross-validation technique (Stone, 1974) is applied to address the challenges of accurately estimating a model's performance on unseen data and to mitigate issues like overfitting. The dataset is partitioned into five subsets (folds), and the model is repeatedly trained and evaluated K times (K = 5). During each iteration, one fold is used as the validation set, while the remaining (K-1) folds are used for training. This process provides a more comprehensive understanding of the model's generalization capability and stability across different data partitions. Table 9 presents the results of 5-fold cross-validation.

Folds	Accuracy	Class	Metrics		
			Precision	Recall	F1-score
1st fold	0.75	Good (Class 0)	0.82	0.77	0.79
		Poor (Class 1)	0.66	0.72	0.69
		Extremely poor (Class 2)	0.78	0.75	0.76
2nd fold	0.74	Good (Class 0)	0.78	0.79	0.79
		Poor (Class 1)	0.68	0.73	0.70
		Extremely poor (Class 2)	0.83	0.57	0.68
3rd fold	0.74	Good (Class 0)	0.74	0.81	0.77
		Poor (Class 1)	0.73	0.76	0.75
		Extremely poor (Class 2)	0.80	0.59	0.68
4th fold	0.79	Good (Class 0)	0.73	0.82	0.77
		Poor (Class 1)	0.85	0.77	0.81
		Extremely poor (Class 2)	0.00	0.00	0.00
5th fold	0.66	Good (Class 0)	0.69	0.57	0.63
		Poor (Class 1)	0.73	0.71	0.72
		Extremely poor (Class 2)	0.30	0.67	0.41

Table 9: Results of the 5-fold cross-validation

As illustrated in Table 9, the application of 5-fold cross-validation yielded valuable insights into the performance of our model. Upon conducting a 5-fold cross-validation on our dataset, we observed consistent and promising outcomes even without handling the imbalanced classification issue. The model demonstrated a high level of stability and generalization, indicating that it is well-suited for making accurate predictions on unseen data. The low variance in performance across the folds suggests that the model is not overfitting to the training data. This is a crucial characteristic, as it implies that the model is effectively learning underlying patterns rather than memorizing the training set.

In conclusion, this section highlights the successful application of the proposed granular computing rule extraction approach to the task of air quality classification. Our comprehensive experimental evaluation showcases the ap-

proach's capability to effectively classify air quality levels based on meteorological variables. The achieved accuracy, precision, and recall on both the training and test datasets affirm the robustness and generalizability of the approach.

5.3. Comparison of the proposed model with machine learning models

In this section, we conduct a comprehensive comparison between the performance of the proposed Granular Computing (GrC) model for rule extraction and that of machine learning (ML) models commonly employed in air quality classification tasks. The purpose is to evaluate the efficacy of the GrC algorithm in extracting meaningful classification rules compared to established ML approaches. We consider benchmark models such as decision tree classifier, random forest classifier, and CatBoost which are known for their competence in classification tasks.

The comparison encompasses various performance metrics, including accuracy, precision, recall, and F1-score, providing a holistic view of the models' effectiveness in distinguishing between different air quality levels. Additionally, we assess the interpretability of the rules generated by each model, a crucial aspect in the context of air quality analysis.

Decision tree classifier is a widely used machine learning algorithm that operates by recursively partitioning the dataset based on features, creating a tree-like structure where each leaf node represents a class. It is known for its simplicity, interpretability, and ability to handle both numerical and categorical data. Decision trees are prone to overfitting, which can be mitigated by ensemble methods like Random Forest.

Random forest is an ensemble learning algorithm that constructs a multitude of decision trees during training and outputs the mode of the classes for classification problems. It excels in reducing overfitting and increasing predictive accuracy by aggregating the results of multiple decision trees. The randomness introduced during the tree-building process enhances robustness and generalizability.

CatBoost, short for Categorical Boosting, is a gradient boosting algorithm

specifically designed to handle categorical features efficiently. It automates the process of encoding categorical variables and incorporates a robust optimization scheme. CatBoost is known for its high performance with minimal hyperparameter tuning, making it suitable for various classification tasks.

Table 10 summarizes the performance metrics of the decision tree classifier, random forest, CatBoost, and our proposed Granular Computing (GrC) algorithm on the air quality classification task. The metrics include accuracy, precision, recall, F1-score, and interpretability. The results aim to provide a comprehensive understanding of each algorithm's strengths and weaknesses in the context of air quality analysis.

Algorithm	Accuracy	Class	Metrics			Interpretability
			Precision	Recall	F1-score	
AirQ-RuleGrCEx	0.79	Good (Class 0)	0.82	0.81	0.81	Yes
		Poor (Class 1)	0.76	0.80	0.78	
		Extremely poor (Class 2)	0.85	0.71	0.77	
Decision Tree	0.79	Good (Class 0)	0.82	0.78	0.80	Yes
		Poor (Class 1)	0.75	0.80	0.78	
		Extremely poor (Class 2)	0.87	0.80	0.84	
Random Forest	0.84	Good (Class 0)	0.87	0.87	0.87	No
		Poor (Class 1)	0.82	0.84	0.83	
		Extremely poor (Class 2)	0.86	0.76	0.81	
CatBoost	0.86	Good (Class 0)	0.86	0.89	0.87	No
		Poor (Class 1)	0.88	0.80	0.84	
		Extremely poor (Class 2)	0.82	0.98	0.89	

Table 10: Comparison of classification algorithms for air quality

As illustrated in Table 10, the proposed model showed competitive performance with other widely used tree-based models. The interpretability aspect is an advantage for the proposed GrC model and decision trees, but random forest and CatBoost are considered to be somewhat interpretable models, but they may be less interpretable compared to individual decision trees and GrC, since random forest is based on ensemble learning method that builds multiple decision trees and combines their predictions, and on the other hand, CatBoost is based on the gradient boosting framework, which builds an ensemble of de-

cision trees sequentially, where each tree corrects the errors of the previous ones, leading to a strong predictive model. That is what makes random forest and CatBoost models less interpretable compared to decision tree and granular computing models.

6. Discussions

In the investigation of air quality classification models, a comprehensive discussion of the outcomes is presented, emphasizing the implications of the proposed granular computing (GrC) approach compared to some other widely used machine learning algorithms. Our comparison with established algorithms such as the decision tree classifier, random forest, and CatBoost reveals that the GrC algorithm delivers competitive performance. This finding underscores the potential of granular computing as a promising methodology in machine learning, particularly for air quality classification tasks. A notable advantage of the GrC algorithm lies in its prioritization of interpretability. The rules extracted by the GrC model offer a clear and human-understandable representation of decision-making processes. This interpretability is invaluable in systems/methods where understanding the rationale behind decisions is as crucial as predictive accuracy. The complexity inherent in air quality dynamics, influenced by various meteorological variables, is effectively addressed by the GrC algorithm. Its capability to create granules and hierarchies allows for the modeling of intricate relationships, contributing to the robustness of our model. A noteworthy aspect is the potential for rule generalization exhibited by the GrC algorithm. By focusing on granular patterns, it may identify common rules applicable across different geographical locations and time periods, enhancing its applicability. While our results are promising, challenges, such as the scalability of the GrC algorithm to larger datasets, need consideration, the ability to take into consideration categorical and numerical variables, and investigating approaches to overcome the overfitting problem. Future work could involve optimizations to improve efficiency and scalability, along with exploration into the adaptation

of GrC for real-time air quality monitoring systems. In conclusion, our study demonstrates the effectiveness of granular computing in air quality classification. The GrC algorithm, with its interpretability and adaptability to complex structures, makes a valuable contribution to the field. As air quality remains a critical environmental concern, our work opens avenues for the development of explainable and effective models for pollution level prediction.

7. Conclusion and prospects

This study presents and proposes a novel method for classifying air quality levels using the granular computing rule extraction method. Our analysis revealed significant insights into the relationship between meteorological variables and $PM_{2.5}$ concentrations, contributing to the fields of machine learning, environmental science, and air quality management. We successfully constructed a decision granular tree using our proposed algorithm, which systematically partitions the dataset into coherent granules. The algorithm effectively extracted classification rules that accurately predict air quality levels based on meteorological attributes. The granular tree model demonstrated high accuracy on both the training and test sets. Moreover, these extracted rules can be employed to classify the level of $PM_{2.5}$ for other datasets in different locations.

The proposed approach in this paper offers several strengths and contributions to the field of air quality classification:

1. Granular computing: The use of granular computing in air quality classification is a unique and promising aspect of this work. By representing data as coherent granules, the model can handle complex and uncertain environmental data more effectively.
2. Rule extraction: The study focuses on extracting rules from granules with zero-entropy value, providing interpretability to the classification model. This is crucial in environmental studies, where transparent and understandable decisions are highly desirable.

3. Efficiency and scalability: The step-by-step approach and granule network construction contribute to an efficient and scalable classification process. The algorithm reduces the computational complexity while still achieving a high classification accuracy.
4. Flexibility: The proposed method is adaptable and can accommodate different types of air quality data, making it applicable to various environmental monitoring scenarios.

In conclusion, this paper presents an innovative method for air quality classification based on granular computing. The proposed approach offers unique advantages in terms of interpretability, efficiency, and adaptability. However, further experimentation, comparison with existing methods, and investigation into robustness and generalization are essential for establishing the approach's practical utility and potential impact on environmental monitoring and public health. In future work, we plan to enhance our method for a better deployment with various types of data (numerical and categorical). We are also looking into creating hybrid techniques that merge the advantages of neural networks and granular computing (GrC-ANN) as introduced in these two studies (Ghiasi et al., 2019, 2022) to take advantage of the potential strength resulting from the interaction, as well as the complementary nature of these techniques for solving classification problems. Additionally, we aim to investigate how a sequential three-way decision approach (a granular computing approach) could be applied to improve air quality forecasting.

CRediT authorship contribution statement

Idriss Jairi: Conceptualization, Methodology, Software, Data curation, Writing - original draft, Formal analysis, Validation. **Sarah Ben-Othman:** Supervision, Writing - review editing, Validation. **Ludivine Canivet:** Supervision, Writing - review editing, Validation. **Hayfa Zgaya-Biau:** Supervision, Methodology, Editing the paper, Validation.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Abdel-Rahman, A. A. (2008). On the atmospheric dispersion and gaussian plume model. In *Proceedings of the 2nd International Conference on Waste Management, Water Pollution, Air Pollution, Indoor Climate, Corfu, Greece*. volume 26.

Abhilash, M., Thakur, A., Gupta, D., & Sreevidya, B. (2018). Time series analysis of air pollution in bengaluru using arima model. In *Ambient Communications and Computer Systems: RACCCS 2017* (pp. 413–426). Springer.

Aggarwal, A., Choudhary, T., & Kumar, P. (2017). A fuzzy interface system for determining air quality index. In *2017 International conference on infocom technologies and unmanned systems (trends and future directions)(ICTUS)* (pp. 786–790). IEEE.

Bargiela, A., & Pedrycz, W. (2006). The roots of granular computing. In *2006 IEEE International Conference on Granular Computing* (pp. 806–809). IEEE.

Bargiela, A., & Pedrycz, W. (2022). Granular computing. In *HANDBOOK ON COMPUTER LEARNING AND INTELLIGENCE: Volume 2: Deep Learning, Intelligent Control and Evolutionary Computation* (pp. 97–132). World Scientific.

Bessagnet, B., Hodzic, A., Vautard, R., Beekmann, M., Cheinet, S., Honoré, C., Liousse, C., & Rouil, L. (2004). Aerosol modeling with chimere—preliminary evaluation at the continental scale. *Atmospheric environment*, *38*, 2803–2817.

Biancofiore, F., Busilacchio, M., Verdecchia, M., Tomassetti, B., Aruffo, E., Bianco, S., Di Tommaso, S., Colangeli, C., Rosatelli, G., & Di Carlo, P. (2017). Recursive neural network model for analysis and forecast of pm10 and pm2. 5. *Atmospheric Pollution Research*, *8*, 652–659.

Binkowski, F. S., & Roselle, S. J. (2003). Models-3 community multiscale air quality (cmaq) model aerosol component 1. model description. *Journal of geophysical research: Atmospheres*, *108*.

Bozdağ, A., Dokuz, Y., & Gökçek, Ö. B. (2020). Spatial prediction of pm10 concentration using machine learning algorithms in ankara, turkey. *Environmental Pollution*, *263*, 114635.

Castelli, M., Clemente, F. M., Popovič, A., Silva, S., & Vanneschi, L. (2020). A machine learning approach to predict air quality in california. *Complexity*, *2020*.

Chelani, A. B., Rao, C. C., Phadke, K., & Hasan, M. (2002). Prediction of sulphur dioxide concentration using artificial neural networks. *Environmental Modelling & Software*, *17*, 159–166.

Chen, S. (2017). Beijing PM2.5 Data. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5JS49.

Corani, G., & Scanagatta, M. (2016). Air pollution prediction via multi-label classification. *Environmental modelling & software*, *80*, 259–264.

Du, S., Li, T., Yang, Y., & Horng, S.-J. (2019). Deep air quality forecasting using hybrid deep learning framework. *IEEE Transactions on Knowledge and Data Engineering*, *33*, 2412–2424.

Eren, B., Aksangür, İ., & Erden, C. (2023). Predicting next hour fine particulate matter (pm2. 5) in the istanbul metropolitan city using deep learning algorithms with time windowing strategy. *Urban Climate*, *48*, 101418.

Ghiasi, B., Noori, R., Sheikhian, H., Zeynolabedin, A., Sun, Y., Jun, C., Hamouda, M., Bateni, S. M., & Abolfathi, S. (2022). Uncertainty quantification of granular computing-neural network model for prediction of pollutant longitudinal dispersion coefficient in aquatic streams. *Scientific reports*, *12*, 4610.

Ghiasi, B., Sheikhian, H., Zeynolabedin, A., & Niksokhan, M. H. (2019). Granular computing–neural network model for prediction of longitudinal dispersion coefficients in rivers. *Water Science and Technology*, *80*, 1880–1892.

Gore, R. W., & Deshpande, D. S. (2017). An approach for classification of health risks based on air quality levels. In *2017 1st International Conference on Intelligent Systems and Information Management (ICISIM)* (pp. 58–61). IEEE.

Hamami, F., & Dahlan, I. A. (2022). Air quality classification in urban environment using machine learning approach. In *IOP Conference Series: Earth and Environmental Science* (p. 012004). IOP Publishing volume 986.

Hamami, F., & Fithriyah, I. (2020). Classification of air pollution levels using artificial neural network. In *2020 International Conference on Information Technology Systems and Innovation (ICITSI)* (pp. 217–220). IEEE.

Haq, M. A. (2022). Smotednn: A novel model for air pollution forecasting and aqi classification. *Computers, Materials & Continua*, *71*.

Jhun, I., Coull, B. A., Schwartz, J., Hubbell, B., & Koutrakis, P. (2015). The impact of weather changes on air quality and health in the united states in 1994–2012. *Environmental research letters*, *10*, 084009.

Kampa, M., & Castanas, E. (2008). Human health effects of air pollution. *Environmental pollution*, *151*, 362–367.

Khamespanah, F., Delavar, M. R., Alinia, H. S., & Zare, M. (2013). Granular computing and dempster–shafer integration in seismic vulnerability assess-

ment. *Intelligent Systems for Crisis Management: Geo-information for Disaster Management (Gi4DM) 2012*, (pp. 147–158).

Kujaroentavon, K., Kiattisin, S., Leelasantitham, A., & Thammaboosadee, S. (2014). Air quality classification in thailand based on decision tree. In *The 7th 2014 Biomedical Engineering International Conference* (pp. 1–5). IEEE.

Kumar, A., & Goyal, P. (2011). Forecasting of air quality in delhi using principal component regression technique. *Atmospheric Pollution Research*, *2*, 436–444.

Kumar, A., & Goyal, P. (2013). Forecasting of air quality index in delhi using neural network based on principal component analysis. *Pure and Applied Geophysics*, *170*, 711–722.

Kumar, U., & Jain, V. (2010). Arima forecasting of ambient air pollutants (o 3, no, no 2 and co). *Stochastic Environmental Research and Risk Assessment*, *24*, 751–760.

Lei, T. M., Ng, S. C., & Siu, S. W. (2023). Application of ann, xgboost, and other ml methods to forecast air quality in macau. *Sustainability*, *15*, 5341.

Liu, H., Wu, H., Lv, X., Ren, Z., Liu, M., Li, Y., & Shi, H. (2019). An intelligent hybrid model for air pollutant concentrations forecasting: Case of beijing in china. *Sustainable Cities and Society*, *47*, 101471.

Ma, J., Yu, Z., Qu, Y., Xu, J., Cao, Y. et al. (2020). Application of the xgboost machine learning method in pm2. 5 prediction: A case study of shanghai. *Aerosol and Air Quality Research*, *20*, 128–138.

Mangayarkarasi, R., Vanmathi, C., Khan, M. Z., Noorwali, A., Jain, R., & Agarwal, P. (2021). Covid19: Forecasting air quality index and particulate matter (pm2. 5). *Computers, Materials & Continua*, *67*.

McHugh, C., Carruthers, D., & Edmunds, H. (1997). Adms–urban: an air quality management system for traffic, domestic and industrial pollution. *International Journal of Environment and Pollution*, *8*, 666–674.

Menut, L., Bessagnet, B., Khvorostyanov, D., Beekmann, M., Blond, N., Colette, A., Coll, I., Curci, G., Foret, G., Hodzic, A. et al. (2013). Chimere 2013: a model for regional atmospheric composition modelling. *Geoscientific model development*, *6*, 981–1028.

Moazami, S., Noori, R., Amiri, B. J., Yeganeh, B., Partani, S., & Safavi, S. (2016). Reliable prediction of carbon monoxide using developed support vector machine. *Atmospheric Pollution Research*, *7*, 412–418.

Niska, H., Hiltunen, T., Karppinen, A., Ruuskanen, J., & Kolehmainen, M. (2004). Evolving the neural network model for forecasting air pollution time series. *Engineering Applications of Artificial Intelligence*, *17*, 159–167.

Noori, R., Ghiasi, B., Sheikhian, H., & Adamowski, J. F. (2017a). Estimation of the dispersion coefficient in natural rivers using a granular computing model. *Journal of Hydraulic Engineering*, *143*, 04017001.

Noori, R., Hoshyaripour, G., Ashrafi, K., & Araabi, B. N. (2010). Uncertainty analysis of developed ann and anfis models in prediction of carbon monoxide daily concentration. *Atmospheric Environment*, *44*, 476–482.

Noori, R., Sheikhian, H., Hooshyaripor, F., Naghikhani, A., Adamowski, J. F., & Ghiasi, B. (2017b). Granular computing for prediction of scour below spillways. *Water Resources Management*, *31*, 313–326.

Osowski, S., & Garanty, K. (2007). Forecasting of the daily meteorological pollution using wavelets and support vector machine. *Engineering Applications of Artificial Intelligence*, *20*, 745–755.

Pawlak, Z. (1982). Rough sets. *International journal of computer & information sciences*, *11*, 341–356.

Pedrycz, W. (2018). *Granular computing: analysis and design of intelligent systems*. CRC press.

Rozehkhani, S. M., & Mahan, F. (2022). Vm consolidation improvement approach using heuristics granular rules in cloud computing environment. *Information Sciences*, *596*, 15–29.

Rozehkhani, S. M., & Mohammadzad, M. (2022). Rule extraction for screening of covid-19 disease using granular computing approach. *Computational and Mathematical Methods in Medicine*, *2022*.

Samadi Alinia, H., & Delavar, M. (2011). Tehran's seismic vulnerability classification using granular computing approach. *Applied Geomatics*, *3*, 229–240.

Saminathan, S., & Malathy, C. (2023). Ensemble-based classification approach for pm2. 5 concentration forecasting using meteorological data. *Frontiers in big Data*, *6*, 1175259.

Sheikhian, H., Delavar, M. R., & Stein, A. (2017). A gis-based multi-criteria seismic vulnerability assessment using the integration of granular computing rule extraction and artificial neural networks. *Transactions in GIS*, *21*, 1237–1259.

Slini, T., Karatzas, K., & Papadopoulos, A. (2002). Regression analysis and urban air quality forecasting: An application for the city of athens. *Global Nest*, *4*, 153–162.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of the royal statistical society: Series B (Methodological)*, *36*, 111–133.

Sugiarto, B., & Sustika, R. (2016). Data classification for air quality on wireless sensor network monitoring system using decision tree algorithm. In *2016 2nd International Conference on Science and Technology-Computer (ICST)* (pp. 172–176). IEEE.

Teologo, A. T., Dadios, E. P., Baldovino, R. G., Neyra, R. Q., & Javel, I. M. (2018). Air quality index (aqi) classification using co and no 2 pollutants: a fuzzy-based approach. In *TENCON 2018-2018 IEEE Region 10 Conference* (pp. 0194–0198). IEEE.

Tsai, Y.-T., Zeng, Y.-R., & Chang, Y.-S. (2018). Air pollution forecasting using rnn with lstm. In *2018 IEEE 16th Intl Conf on Dependable, Autonomic and Secure Computing, 16th Intl Conf on Pervasive Intelligence and Computing, 4th Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress (DASC/PiCom/DataCom/CyberSciTech)* (pp. 1074–1079). IEEE.

Wang, J., & Song, G. (2018). A deep spatial-temporal ensemble model for air quality prediction. *Neurocomputing*, *314*, 198–206.

WHO (2022). World Health Organization ambient (outdoor) air pollution. URL: `https://www.who.int/news-room/fact-sheets/detail/ambient-(outdoor)-air-quality-and-health`.

Yao, J., & Yao, Y. (2002). Induction of classification rules by granular computing. In *Rough Sets and Current Trends in Computing: Third International Conference, RSCTC 2002 Malvern, PA, USA, October 14–16, 2002 Proceedings 3* (pp. 331–338). Springer.

Yao, Y. (2004). A partition model of granular computing. In *Transactions on Rough Sets I: James F. Peters-Andrzej Skowron, Editors-in-Chief* (pp. 232–253). Springer.

Zadeh, L. A. (1965). Fuzzy sets. *Information and control*, *8*, 338–353.

Zadeh, L. A. (1979). Fuzzy sets and information granularity. *Fuzzy sets, fuzzy logic, and fuzzy systems: selected papers*, (pp. 433–448).

Zadeh, L. A. (1997). Toward a theory of fuzzy information granulation and its centrality in human reasoning and fuzzy logic. *Fuzzy sets and systems*, *90*, 111–127.

Zaini, N., Ean, L. W., Ahmed, A. N., Abdul Malek, M., & Chow, M. F. (2022). Pm2. 5 forecasting for an urban area based on deep learning and decomposition method. *Scientific Reports*, *12*, 17565.

Zhang, Z., & Chen, Q. (2007). Comparison of the eulerian and lagrangian methods for predicting particle transport in enclosed spaces. *Atmospheric environment*, *41*, 5236–5248.

Zhao, X., Zhang, R., Wu, J.-L., & Chang, P.-C. (2018). A deep recurrent neural network for air quality classification. *J. Inf. Hiding Multim. Signal Process.*, *9*, 346–354.

Zhao, Y., Hasan, Y. A. et al. (2013). Comparison of three classification algorithms for predicting pm2. 5 in hong kong rural area. *Journal of Asian Scientific Research*, *3*, 715–728.