*minerals*

MDPI

*Review*

# Big Data and AI-Enabled Construction of a Novel Gemstone Database: Challenges, Methodologies, and Future Perspectives

**Yu Zhang and Guanghai Shi ***

School of Gemmology, China University of Geosciences (Beijing), Beijing 100083, China; zhangyu@cugb.edu.cn
* Correspondence: shigh@cugb.edu.cn

**Abstract**

Gemstone samples, as objects of study in gemology, carry rich geological information and cultural value, playing an irreplaceable role in teaching, research, and public science communication. In the current age of big data, machine learning and artificial intelligence techniques based on gemstone databases have emerged as a cutting-edge area of gemology. However, traditional gemstone databases have three major limitations: an absence of standardized data schemas, incomplete core datasets (e.g., records of synthetic and treated gemstones and inclusion characteristics), and poor data interoperability. These deficiencies hinder the application of advanced technologies, such as machine learning (ML) and AI techniques. This paper reviews gemstone data and applications, as well as existing gem-related sample databases, and proposes a framework for a new gemstone database based on standardization (FAIR principles), integration (blockchain technology), and dynamism (real-time updates). This framework could transform the gemstone industry, shifting it from "experience-driven" to "data-driven" practices. Powered by big data technology, this novel database will revolutionize gemological research, jewelry authentication, market transactions, and educational outreach, fostering innovation in academic research and practical applications.

**Keywords:** big data; gemstones; database; artificial intelligence; machine learning; gemstone identification

check for updates

Academic Editors: Behnam Sadeghi, Yongzhang Zhou, Hui Yang and Xiaohui Ji

Received: 12 September 2025
Revised: 27 October 2025
Accepted: 29 October 2025
Published: 31 October 2025

**Citation:** Zhang, Y.; Shi, G. Big Data and AI-Enabled Construction of a Novel Gemstone Database: Challenges, Methodologies, and Future Perspectives. *Minerals* **2025**, *15*, 1149. https://doi.org/10.3390/min15111149

**Copyright:** © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

## 1. Introduction

With the rapid development of global information technology, big data has become a significant driver of innovation in various fields [1–3], including the traditional gemstone industry, which is undergoing unprecedented changes. Gemology, as a science, can be traced back to ancient humans' exploration and research of precious minerals. The advent of modern technology has led to the evolution of gemology into a multidisciplinary field of study, encompassing physics, chemistry, earth sciences, material science, archaeology, art, commercial trade and other fields.

Gemstone samples serve as objects of study in gemology, carrying rich geological information and cultural value. Most gemstones are beautiful, rare, durable minerals that can be used for decoration. Unlike natural mineral samples, gemstone samples include not only natural gemstones, but also synthetic and treated gemstones, all of which circulate on the market as gemstones [4–6]. In education, gemstone specimens help students observe the physical properties and crystal structures of minerals, learn to identify natural gemstone varieties and their origins, and distinguish between various synthetic and treated gemstones [5]. In scientific research, gemstone specimens provide crucial physical evidence

for studies on ore deposit formation, geological evolution, and the mechanisms behind gemstone coloration [7,8]. The vibrant and colorful gemstone mineral samples are a great way to get the public interested in Earth sciences. They can also help people learn more about gemstones and improve their scientific literacy.

Gemology is a data-intensive discipline. Traditional gemstone databases are affiliated with various universities, museum, institutions, and jewelry laboratories. These databases record the basic physical and chemical properties, optical characteristics, and limited spectral data of the gemstones they own and have tested. These databases face several challenges, including single-source data, limited data scale, long update cycles, and a lack of analytical capabilities. Additionally, inconsistent data is difficult to integrate due to a lack of standardized norms and standards among institutions. Gaps in mineralogical and geochemical data for these gemstones also require urgent attention. Some mineral databases, such as Mindat.org and RRUFF, contain gemstone data as well. However, they lack key gemological information, such as data on synthetic gemstones, gemstone enhancements, and the causes of gemstone colors. Furthermore, traditional manual operations and limited data processing capabilities are insufficient to meet the needs of the gemstone industry.

Big data technology provides immense computational and storage capabilities, enabling the efficient processing and analysis of large volumes of complex, specialized data [1]. These capabilities form the foundation for the implementation of machine learning and artificial intelligence. Consequently, gemological information management and application have entered a new phase. Unlike traditional research paradigms, which emphasize causal relationships, big data research methods rely on correlations to identify issues [2]. In addition to answering existing scientific questions, applying big data to research can unlock the potential of recorded data and aid in exploring previously undiscovered phenomena [9]. In gemology, big data research has primarily focused on origin tracing, gemstone identification, gemstone grading and estimation of mineralization. This research has yielded notable achievements, such as identifying the origins of emeralds [10] and jadeite jade [11], developing intelligent gemstone identification systems (such as the Gemtelligence system from the Gübelin Laboratory) [12], emerald gemstone grading [13], and estimation of emerald mineralization probability [14].

The introduction of big data technology has presented a significant opportunity for the digital transformation of the gemstone industry. However, big data research in the field of gemology faces unprecedented challenges due to various factors, such as the scarcity of global gemstone resources, the high cost of gemstone specimens, and the fact that core data is treated as commercial secrets and rarely shared.

In light of this, we propose an initiative to establish a new gemological database with appropriate data standards to integrate heterogeneous data. This database will enable the efficient storage and management of gemstone information, facilitate real-time updates and intelligent analysis, and overcome the limitations of traditional databases. The database will provide stronger support for gemstone identification research, origin tracing, market analysis, heritage science, and other related fields.

## 2. Application of Gemstone Databases in Gemstone Research

Gemstone research data includes information on physical and chemical properties, images (including characteristics of the internal and external structure), spectra, compositions, and geochemistry. Gemstone databases play a multidimensional supporting role in research, and their core value can be summarized in four ways. Each aspect achieves a leap in capability through big data technology.

### *2.1. Gemstone Identification and Testing*

Gemstone identification involves determining the species of a gemstone, as well as identifying synthetic gemstones and enhanced treatment methods. In recent years, advances in laboratory synthesis technology and optimized treatment technology have provided the jewelry market with more product choices. However, due to significant differences in market value, determining whether a gemstone is synthetic or has undergone treatment has become a top priority.

Gemstone identification and testing methods essentially involve comparing data from gemstone samples with known data. In recent years, non-destructive analytical techniques, such as ultraviolet-visible (UV-Vis) spectroscopy, Raman spectroscopy, photoluminescence (PL) spectroscopy, Fourier transform infrared (FTIR) spectroscopy, and cathodoluminescence (CL), have been widely applied in gemstone research. Characterized by their non-destructive nature, high sensitivity, and high resolution, these techniques have become essential tools in gemological research. Comparative studies of datasets generated by these methods are common in gemological research.

For example, Culka et al. (2019) [15] used a portable, sequentially shifted excitation Raman spectrometer to obtain Raman spectra of cut gemstones. They then loaded these spectra into the free CrystalSleuth program and retrieved them from the online RRUFF database to identify the gemstones. Zhang et al. (2024) [5] conducted a comparative analysis of natural aquamarine from the Altay region of Xinjiang, China and hydrothermally synthetic aquamarine available on the Chinese market. They found that infrared absorption bands at 3313 and 4060 $cm^{-1}$, which are caused by N-H bond vibrations, were only present in the hydrothermally synthetic aquamarine and can be used to distinguish natural aquamarine from hydrothermally synthetic aquamarine (Figure 1).

It is worth noting that the aforementioned analytical techniques are often used in combination to examine gemstones more comprehensively. For instance, in diamond-related research, the type of diamond can be identified, as well as the method used to synthesize it and any subsequent treatments.

The difference between synthetic and natural diamonds is based on variations in impurities and their arrangement, which result from their different growth settings [6,16]. Two classifications of diamonds are commonly recognized: nitrogen-containing and nitrogen-free. Nitrogen-containing diamonds, designated as Type Ia and Type Ib, contain aggregated nitrogen impurities or isolated nitrogen atoms, respectively. In contrast, nitrogen-free diamonds, classified as Type IIa and Type IIb, lack nitrogen and contain boron impurities [16–18]. The presence of nitrogen-related (e.g., 1175, 1282, 1130, and 1344 $cm^{-1}$) and boron-related (e.g., 2930, 2803, and 2458 $cm^{-1}$) infrared (IR) absorption peaks, as well as optical defects (e.g., N3 centers), as determined by UV-visible and photoluminescence (PL) spectroscopy, can help identify the type of diamond [6]. This is very helpful for diamond identification, as over 98% of natural diamonds are Type Ia, most HPHT diamonds are Type Ib, and CVD synthetic diamonds are generally Type IIa. Additionally, diamond-view fluorescence imaging technology is widely used to identify synthetic diamonds, including HPHT and natural diamonds, as well as CVD diamonds, which exhibit distinct emission patterns. Synthetic diamonds undergo post-growth treatments, including high-pressure, high-temperature (HPHT) processing or low-pressure, high-temperature (LPHT) annealing, to improve quality [19–22]. These treatments cause changes in defect concentration and spectral characteristics, including infrared, ultraviolet-visible (UV-Vis) absorption, Raman, and PL. These spectral data differences can serve as important evidence for identifying treated diamonds.

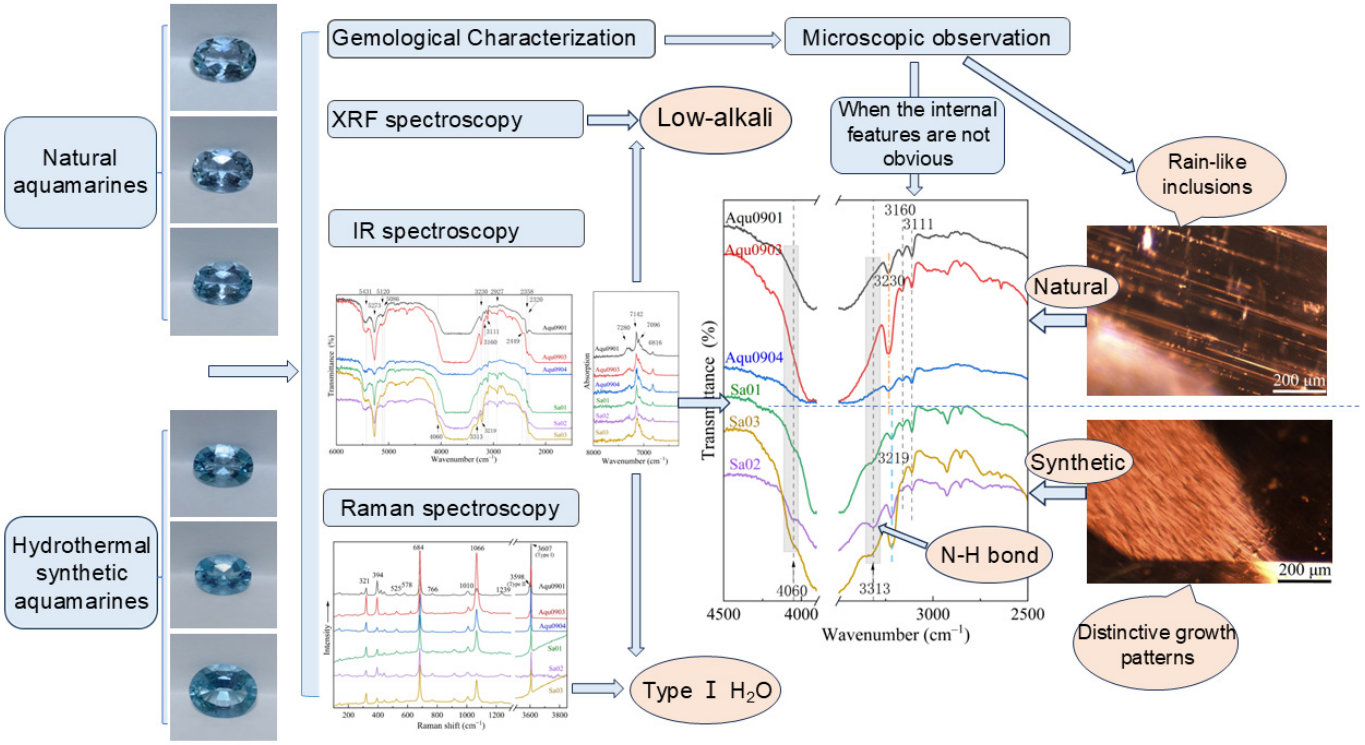



**Figure 1.** A comparative analysis of natural aquamarine from the Altay region of Xinjiang, China and hydrothermally synthesized aquamarine available on the Chinese market was conducted using microscopic observation and spectroscopic analysis (XRF, IR, and Raman). The results indicate that the infrared absorption bands at 3313 and 4060 $cm^{-1}$ (which are attributed to N-H bond vibrations) are exclusively present in the hydrothermally synthetic aquamarine. This enables the differentiation between the natural and synthetic aquamarines (cited in [5]).

Accurate identification requires the combined use of multiple technologies and an in-depth analysis of gemstone data. This lays a solid foundation for the subsequent construction and maintenance of databases.

### *2.2. The Formation of Gems and Their Geographical Origins*

The formation and geographical origins of gems constitute another major area of study in gemology. The distinctive geological environments in which many gems are found offer invaluable insights into the evolutionary history of tectonic structures. In contrast to minerals, the geographical provenance of gems is also of great significance, often associated with substantial economic interests.

A thorough examination of the gemstone's fundamental characteristics, including its refractive index, specific gravity, hue, and transparency, can assist in ascertaining its origin. These figures are rooted in in-depth statistical assessments and represent a body of empirical knowledge. Inclusions, which are foreign substances captured during gem formation, are often key indicators for determining gem origin [23,24]. The types and quantities of inclusions in emeralds vary depending on their origin. While certain inclusions are relatively common worldwide, specific combinations can still provide valuable clues about the origin of a particular emerald [23,25].

The spectroscopic characteristics and chemical composition of gemstones are crucial for tracing their origin. Due to the substantial differences in market value attributed to their various origins, emeralds, as precious gemstones, have garnered significant attention from gemologists in origin research. Many researchers have obtained UV-Vis-NIR spectra, IR spectra, Raman spectra, and LA-ICP-MS data from emerald samples in multiple countries and classified and compared these data statistically. These research findings are based on

datasets obtained by researchers, resulting in binary/ternary identification diagrams and identification flowcharts (Figure 2).

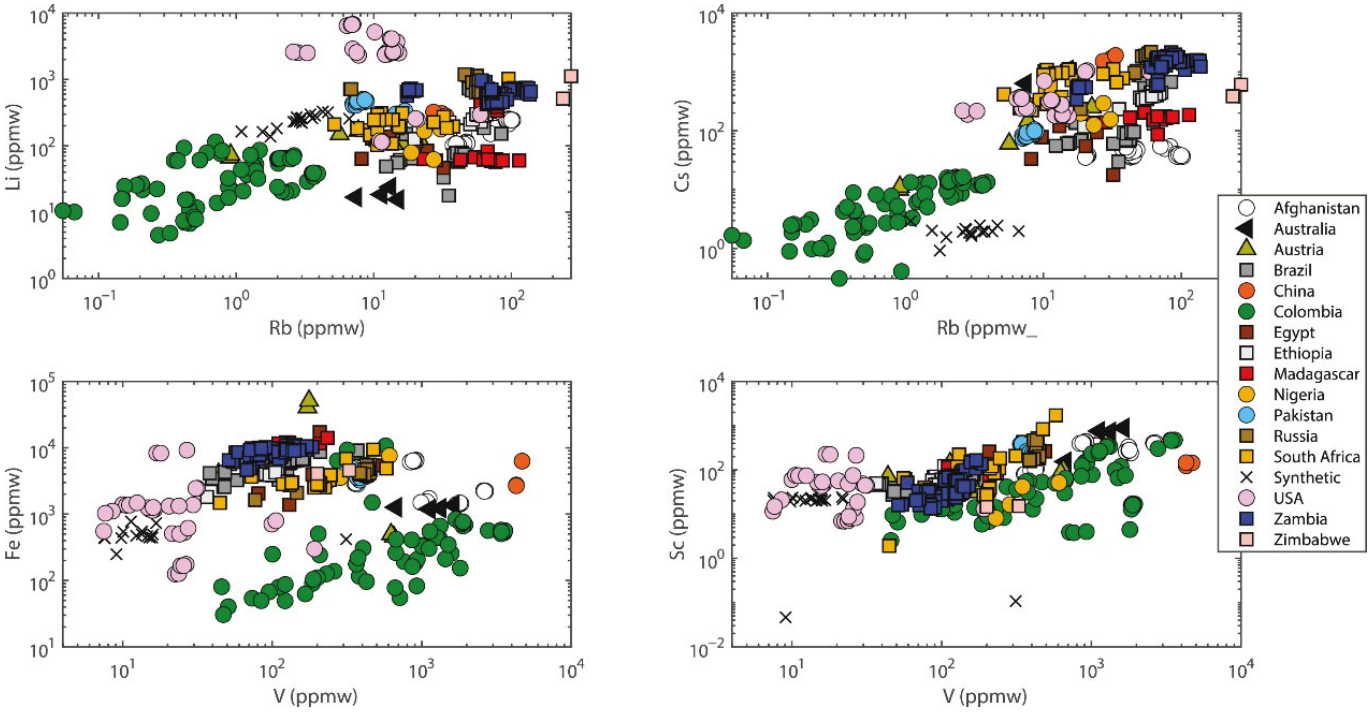



**Figure 2.** Plots of Rb concentrations versus Li and Cs, as well as V concentrations versus Fe and scandium Sc, for emerald origin determination (cited in [10]).

In recent years, research findings on emeralds from new sources, such as Ethiopia, Madagascar, Austria, Ukraine, and others, have gradually emerged [26–29]. Meanwhile, with the development of trace element testing methods, new data on emeralds from traditional mining regions, such as Egypt, Pakistan, India, Brazil, Afghanistan, and Colombia, has also been collected [30–34]. The increasing availability of chemical composition data enhances the feasibility and reliability of using it for origin identification. However, as the dataset grows, significant overlap has emerged in traditional binary/ternary identification charts, necessitating improvements to traditional statistical methods and the introduction of technologies such as machine learning and artificial intelligence.

### *2.3. Mechanism of Gemstone Color Formation*

The study of gemstone color formation is an important topic in gemology. An in-depth analysis of color mechanisms can improve the accuracy of gemstone valuation and provide a scientific basis for gemstone enhancement, as well as synthetic gemstone research and development.

Spectroscopic techniques have played a crucial role in investigating the color formation mechanisms of gemstones. Significant progress has been made over the past decade in studying the color formation of jadeite jade using analytical technologies such as ultraviolet-visible absorption spectroscopy (UV-Vis), electron probe microanalysis (EPMA), laser ablation inductively coupled plasma mass spectrometry (LA-ICP-MS), and electron paramagnetic resonance spectroscopy (EPR) [8].

In addition to the natural coloration mechanisms of gemstones, the coloration mechanisms of gemstone color enhancement processes and their identification are also popular research topics.

The most common method currently used in the market to enhance the color of sapphire involves diffusing beryllium (Be) into the crystal at temperatures exceeding 1800 °C [4,35]. The addition of beryllium to the sapphire crystal lattice creates discrete energy levels by transferring electrons to $Be^{2+}$ hole acceptors, i.e., inter-valence charge transfer [36], which alters the sapphire's absorption of visible light, thereby changing its color [35]. Beryllium diffusion treatment technology has been used since at least 2002 to alter the color of corundum, primarily to enhance the color saturation of light-colored sapphires and transform them into vibrant yellow or orange hues. In 2003, the GIA laboratory revealed that some padparadscha sapphires (pink-orange hues of sapphire) on the market had abnormal coloration caused by beryllium diffusion treatment, sparking widespread industry attention [36]. Furthermore, from 2023 to 2024, several international gemological laboratories observed that deep blue sapphires could change in saturation through beryllium diffusion treatment, producing a shift towards a more vivid blue hue.

Beryllium (Be), as a lightweight alkali metal with a concentration of only a few to several dozen ppm in treated sapphire, makes it impossible for traditional compositional testing methods such as XRF and electron probe spectroscopy to determine whether corundum has undergone beryllium diffusion treatment. Currently, jewelry testing institutions use large-scale instruments, such as laser ablation inductively coupled plasma mass spectrometry (LA-ICP-MS) and laser-induced breakdown spectroscopy (LIBS), to determine the beryllium content in corundum. The surface beryllium concentration of natural sapphires ranges from 1.5 to 5 ppm, whereas that of beryllium-diffused sapphires ranges from 10 to 35 ppm. If the beryllium content is below 10 ppm during testing, multiple-point testing should be conducted for a comprehensive assessment to confirm whether the gemstone has undergone beryllium diffusion treatment. However, these instruments are expensive and require high-quality reference samples and specialist operators.

The replacement of cation sites in sapphire by beryllium (Be) leads to changes in Raman spectral characteristics, such as changes in bandwidth and shift. Based on this principle, Chang et al. (2016) conducted a comparative study of the Raman spectra of natural sapphire and beryllium-diffused sapphire [4]. Due to the microstructural and compositional inhomogeneity of the samples, the researchers employed a wide-area illumination (WAI) scheme capable of covering a large sample area (28.3 $mm^2$, with a laser illumination diameter of 6 mm) in order to collect representative Raman spectra of sapphire. They then used principal component analysis (PCA) to represent spectral features in a score domain. Finally, they employed linear discriminant analysis (LDA) and k-nearest neighbor (k-NN) algorithms to classify the samples as natural sapphire or beryllium-diffused sapphire. The cross-validation classification error was 7.3%. This method would be more practical with a statistically valid dataset including a large number of diverse sapphire samples.

### *2.4. Archaeological Research*

Research at the intersection of gemology and archaeology is often inextricably tied to ancient cultures and trade networks.

Butini et al. (2018) [37] studied the sapphires embedded in a Roman imperial necklace, which dates back to the 3rd century AD and was unearthed near in Colonna, near Rome, in 2011. Based on the results of trace element analysis and a comparison with known sapphire origin datasets, they speculated that the three non-basaltic sapphires may have originated in Sri Lanka, and the four basaltic sapphires were sourced from Thailand and Cambodia [38,39]. The researchers also compared these gems with a large gold bracelet set with sapphires, emeralds, and blue glass housed in the J. Paul Getty Museum (collection number 83 AM 227.2). The bracelet dates back to the 250–400 CE period. The sapphires on the bracelet were sourced from Sri Lankan deposits. These sapphires from South and

Southeast Asia reached Rome via lengthy trade routes, as documented by contemporary trade and contact records. This information provides important clues for understanding trade routes and the luxury goods market during the Roman Empire.

## 3. Applications of Big Data and Machine Learning

With the advent of the big data era, multivariate statistics and machine learning methods have emerged as new tools for gem research in the field of gemology. Machine learning methods can quickly process large amounts of gemstone sample data by building big data models, which significantly improves identification speed and accuracy. The core of machine learning lies in simulating human learning and reasoning capabilities through computer programs, which combine data from various sources (such as images, spectra, and chemical composition information) to uncover patterns and make predictions or classifications. This multidimensional analytical approach helps reveal hidden correlations within the data, thereby enhancing the reliability of the results. Compared to traditional methods, the analytical process of artificial intelligence (AI) system is objective and controllable, minimizing the influence of human interference, and the results are reproducible. The exponential accumulation of datasets combined with machine learning technology has significantly advanced the field of gemology by improving the efficiency and accuracy of data-driven decision-making in scientific research.

### *3.1. Application of Computer Vision Technology in Gemstone Identification*

In recent years, computer vision technology has been widely applied in the field of gemstone identification. Machine learning techniques in this field typically involve feature extraction methods followed by basic classification algorithms, which have achieved encouraging results in automating various gemological tasks, such as classifying gems by type and shape, and even more complex tasks like gem grading. These research methodologies are generally divided into four steps: (1) image acquisition, (2) background extraction, (3) feature extraction, and (4) classification, and all provide complete software and hardware solutions [13].

Freire et al. (2022) [40] proposed a deep learning based pretrained model for classifying 87 types of gemstones. Using data augmentation techniques, they artificially expanded the scale of the training dataset and achieved an accuracy rate of 72% with the Inception V3 model. Chakraborty et al. (2024) [41] used a Dense Convolutional Network (DenseNet) [42] for automatic gemstone classification (Figure 3). They constructed a model to classify 87 types of gemstones using a dataset of 3200 gemstone images [41]. The DenseNet 121 and DenseNet 169 models achieved accuracy rates of 76% and 79%, respectively. The DenseNet169 transfer learning (TL) model outperformed existing models developed by others in terms of accuracy, loss function, recall, precision, and F1 score [40,43,44].

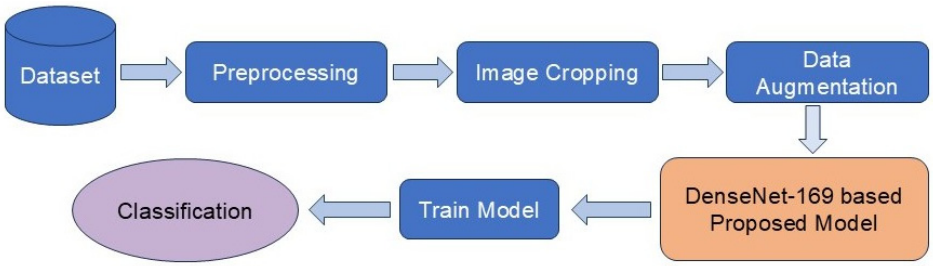



**Figure 3.** The Systematic Workflow Using pre-trained the DenseNet 169 Model for Gemstone Classification.

Currently, gemstone grading is a manual process performed by gemologists. A common method involves the use of reference stones, where available reference stones are

visually inspected by experts to determine which one most closely resembles the gemstone being examined. However, this method is highly subjective, as different experts may produce different grading results.

To address this issue, researchers have conducted gemstone grading studies using image processing techniques. For example, Sinkevicius et al. (2015) [45] developed a system that can identify amber gemstones based on their color and shape. Wang et al. (2016) [46] proposed a system that grades opals using imaging and statistical learning. Pena et al. (2022) [13] applied machine learning to achieve intelligent grading of emerald gemstones. They proposed a framework based on image acquisition protocols, feature extraction, and machine learning. The entire process, from gemstone image acquisition to final gemstone classification, is completed simply by placing the gemstone into the created chamber for image acquisition. This framework combines machine learning methods with image processing techniques for emerald grading. The framework achieved an accuracy rate of 98% (correctly classified gems), outperforming deep learning methods. Additionally, they created and released a public dataset of emeralds for replication and comparison containing 192 images of emeralds and their extracted, preprocessed features.

### *3.2. Intelligent Upgrading of Identification Technology*

The chemical composition and molecular structure of gemstones can be effectively detected using spectroscopy techniques, such as Raman and infrared spectroscopy. Combined with microscopic observation techniques, they can reveal patterns of change in gemstone surface and interior features. However, these techniques present a high barrier to entry for inexperienced operators and cannot directly address the challenges of processing large-scale sample data. Additionally, gemstones treated with new techniques, such as beryllium-diffused sapphires, often appear in large quantities on the market before raising suspicions in gemological laboratories. It is highly likely that sellers sold such gemstones as natural gemstones in the market before gemological laboratories successfully identified the treatment method. Laboratory identification of new methods often lags behind market trends. Nevertheless, the introduction of big data and machine learning has led to intelligent upgrades in gemological techniques.

Khalilian et al. (2024) [47] used laser-induced breakdown spectroscopy (LIBS) and a convolutional neural network long short-term memory (CNN-LSTM) deep learning algorithm to classify different types of jewelry rocks, including azurite, turquoise, calcite, and agate, from the Shahr-e Sukhteh (the Burnt City) in Iran. These rocks span various historical periods and styles. The study's statistical analysis drew from a total of 150 spectra, including 59 turquoises, 46 azurites, 43 agates, and 20 calcites. The CNN-LSTM architecture combines convolutional neural network layers to extract features from the input data and uses long short-term memory networks to forecast sequences. This study provides the first hierarchical interpretation of convolutional LSTM effectiveness, enabling the adaptive acquisition of LIBS features and quantitative data on the main chemical elements in jewelry rocks. The results suggest that combining LIBS with deep learning algorithms can significantly improve the classification of different jewelry samples.

Bendinelli et al. (2024) [12] first proposed a deep learning-based method called GEMTELLIGENCE that automatically determines the origin (OD) of gemstones and detects their treatment (TD) (Figure 4). Developed in collaboration with the Gübelin Gem Laboratory and CSEM, GEMTELLIGENCE can identify the country of origin of rubies, sapphires, and emeralds and detect heat treatment in rubies and sapphires. GEMTELLIGENCE uses a neural network with convolutional and attention mechanisms. The proposed method's primary innovation lies in its multimodal design, which integrates measurement results from four distinct data sources: FTIR (fourier-transform infrared spectroscopy) and UV-Vis

(ultraviolet–visible–near-infrared spectroscopy) for spectral analysis and LA-ICP-MS (laser ablation inductively coupled plasma mass spectrometry) and ED-XRF (energy dispersive X-ray fluorescence) for chemical analysis.

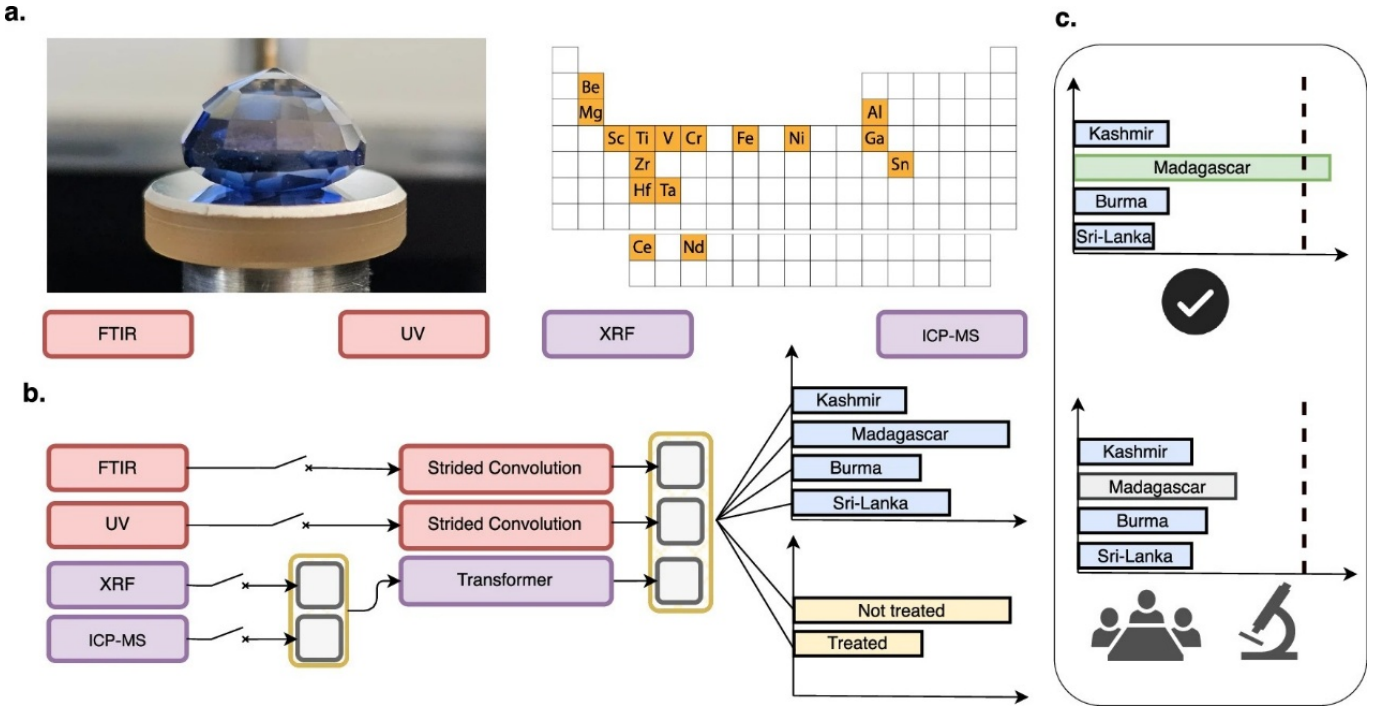



**Figure 4.** (**a**) Measurements from four distinct data sources can be processed by GEMTELLIGENCE: FTIR, UV spectroscopy, ICP-MS, and XRF are used for elemental and spectral analysis. (**b**) The system's prediction of a gemstone's origin or history of heat treatment is accurate and rapid, with results comparable to those of human experts. Not all data types are required for inference; missing sources can be masked as illustrated by the switch symbols in the figure. (**c**) The stone prediction can be confidently accepted if the maximum probability exceeds a predefined threshold (as shown in the top panel, marked with a checkmark). The stone's output should be discarded and its analysis continued using standard methods (e.g., microscopy and expert analysis) if the maximum probability falls below the threshold (bottom panel). The number of stones that can be processed automatically is balanced with the model's accuracy, determined by the value of the threshold selected during the confidence-thresholding phase. (cited in [12]).

The system uses a deep learning architecture trained on tens of thousands of datasets collected by the Gübelin Gem Lab over decades. These datasets include customer and reference gemstone data. Reference samples are defined as gemstones of known origin, which are typically collected by authorized personnel in or near gemstone mining areas, and are accompanied by precise origin records to confirm their provenance. A supervised learning method was employed, whereby the software was fed available analytical data and final results. This method is optimized specifically for multimodal analysis data from various detection devices, enhancing prediction accuracy by leveraging correlations between different data modalities.

Consider sapphire origin identification and heat treatment analysis, for example. With a training set of over 5500 sapphire measurement records, GEMTELLIGENCE was found to achieve outstanding performance by utilizing low-cost data sources. This reduces reliance on expensive analytical methods, such as ICP-MS. Its results are well-calibrated and provide accurate predictions with high confidence across a large number of test samples. At the same time, GEMTELLIGENCE can identify information unknown to the laboratory. For instance, when determining the heat treatment status of sapphires, human experts

primarily focus on the FTIR spectral region between 2500 and 4000 $cm^{-1}$. However, the algorithm appears to retrieve information from regions above 4000 wavenumbers as well. These regions, largely overlooked by human experts, appear to be precisely what enables Gemtelligence to outperform human experts.

However, this type of research has some limitations. For instance, GEMTELLIGENCE's research on sapphires focuses exclusively on metamorphic sapphires. These sapphires, which originate from metamorphic deposits in Kashmir, Myanmar/Burma, Sri Lanka, and Madagascar, account for over 90% of the high-quality sapphires on the market. While these sapphires dominate the market in terms of value, they do not represent all gemstone categories. This limitation primarily stems from the lack of large datasets for non-metamorphic sapphires. Additionally, GEMTELLIGENCE's TD and OD prediction capabilities are validated using gemstone identification results from the Gübelin Gem Lab. While these results are highly accurate, they may not always perfectly reflect the gemstone's true origin. Both limitations could be addressed by introducing refined datasets containing more real-world data.

## 4. Existing Databases and Datasets Related to Gemstones

The field of gemology has accumulated a vast amount of structured and unstructured data over its long history of research. Recognizing the value of these data, jewelry testing institutions or laboratories, geological institutions, universities, and scientists worldwide have established numerous national and international databases (Table 1). While all well-known commercial jewelry testing institutions have their own databases (such as those operated by Gübelin, SSEF, GIA, and GRS), most are not publicly accessible due to commercial confidentiality. Some insight into these databases can be gained through the institutions' published news and research findings. The Hyperion and Sherlock Advanced Gemstone Inclusion Database, created by Lotus Gemology Laboratory, is a specialized database system dedicated to the research and identification of gemstone internal inclusions [48] (Figure 5). It is open to the public and allows users to search for inclusion information based on gemstone type, variety, or origin. The database contains an extensive collection of high-resolution images of internal inclusions from a wide variety of gemstones of different origins, as well as diverse enhanced gemstones. Each image includes detailed information about the gemstone variety, origin, inclusion type (e.g., mineral crystals, fluid inclusions resembling fingerprints, growth lines), and characteristic descriptions. These informations are critical for gemological research and identification.

Some universities also maintain gemstone databases. One example is the School of Gemmology at China University of Geosciences (Beijing). While the databases contain a variety of gemstone samples, access is restricted to internal management and teaching purposes only, and they offer basic search and information retrieval functions (Figure 6).

**Table 1.** List of some existing publicly accessible gemstone-related databases.

| Database Name | Distinctive Functionality | URL |
| --- | --- | --- |
| The Hyperion and Sherlock Gemstone Inclusion Database (Lotus Gemology Laboratory) | A specialized database system is dedicated to the research and identification of gemstone internal inclusions. It is open to the public and allows users to search for inclusion information based on gemstone type, variety, or origin. The database contains an extensive collection of high-resolution images of internal inclusions from a wide variety of gemstones of different origins. | https://lotusgemology.com/index.php/en/resources/hyperion-inclusion-repository, accessed on 27 October 2025 |

**Table 1.** *Cont.*

| Database Name | Distinctive Functionality | URL |
| --- | --- | --- |
| Mindat.org | Mindat.org is the world's leading authority on minerals and their localities, deposits, and mines worldwide. The website provides access to information on mineral chemical formulas, theoretical chemical compositions, occurrences, typical localities, naming origins, crystallographic characteristics, powder X-ray diffraction peaks, physical properties, optical mineralogical features, mineral classifications, related literature, and resource links for natural minerals, including gemstones. It places particular emphasis on providing detailed latitude and longitude coordinates and maps for typical mineral localities. | https://www.mindat.org/, accessed on 27 October 2025 |
| RRUFF | The RRUFF Project is a NASA Mars Mission-derived database that contains an integrated collection of Raman spectra, IR spectra and X-ray diffraction for minerals. It is one of the most authoritative mineral structure databases in the mineralogical community and contains data on natural and synthetic gemstones. | http://rruff.info/, accessed on 27 October 2025 |
| National Infrastructure of Mineral, Rock and Fossil for Science and Technology (NIMRF) | This platform integrates rock, mineral, ore, fossil, and rock core specimen resources. Its systematic mineralogy database query system provides fundamental data on 3600 minerals, including their classification, chemical composition, crystal structure, crystalline morphology, and physical and chemical properties, as well as their genesis and occurrence. It also includes images of select minerals, along with 3D crystal morphology and structural diagrams. There is also a dedicated section on precious gemstones and jade. | http://www.nimrf.net.cn/, accessed on 27 October 2025 |

Currently, gemological research primarily references and utilizes publicly shared databases related to mineralogy and petrology. Among these, Mindat.org is one of the most comprehensive mineral databases, containing information on over 1 million mineral species and their locations, covering more than 300,000 sites, and supporting a significant amount of scientific research. This database places particular emphasis on mineral location information, making it especially useful for mineral specimen collectors. Xiaogang Ma et al. developed the Mindat Open Data Service, the OpenMindat project, achieving two key advancements: (1) improving data quality and (2) creating a data-sharing platform with machine interfaces for querying and accessing data [49].

Another vital resource is the RRUFF database, which is frequently consulted in gemstone research. The RRUFF database is derived from NASA's Mars missions. It comprehensively catalogs mineral Raman spectra, infrared spectra, and X-ray diffraction data [50]. It is one of the most authoritative mineral spectral databases in the field and contains information on natural and synthetic gemstones.

In terms of database resources in China, the National Infrastructure of Mineral, Rock and Fossil for Science and Technology, which is part of the National Science and Technology Infrastructure Platform under the Ministry of Science and Technology, has integrated information on 290,000 mineral and fossil specimens. Its systematic mineralogy database provides basic information on 3600 minerals, including their classification, chemical composition, crystal structure, crystalline morphology, and physical and chemical properties, as well as their origin and occurrence. The database also includes images of minerals, as well as three-dimensional crystal morphology and structural diagrams. The website offers a quick search function for mineralogical data using 16 query criteria, including mineral names in Chinese and English, crystal systems and classes, chemical composition, and color, enabling public access to these resources. The National Rock and Mineral Fossil Specimen Repository also offers a variety of educational resources, including special topics

on precious gemstones and mineral ornamental stones, as well as a series of video presentations. This special topic information includes gemstone properties, color, quality and craftsmanship evaluations, identification treatments, synthesis and identification, differentiation from similar varieties and counterfeits, origin, market, maintenance, and appreciation of famous and premium jewelry and gemstone pieces.

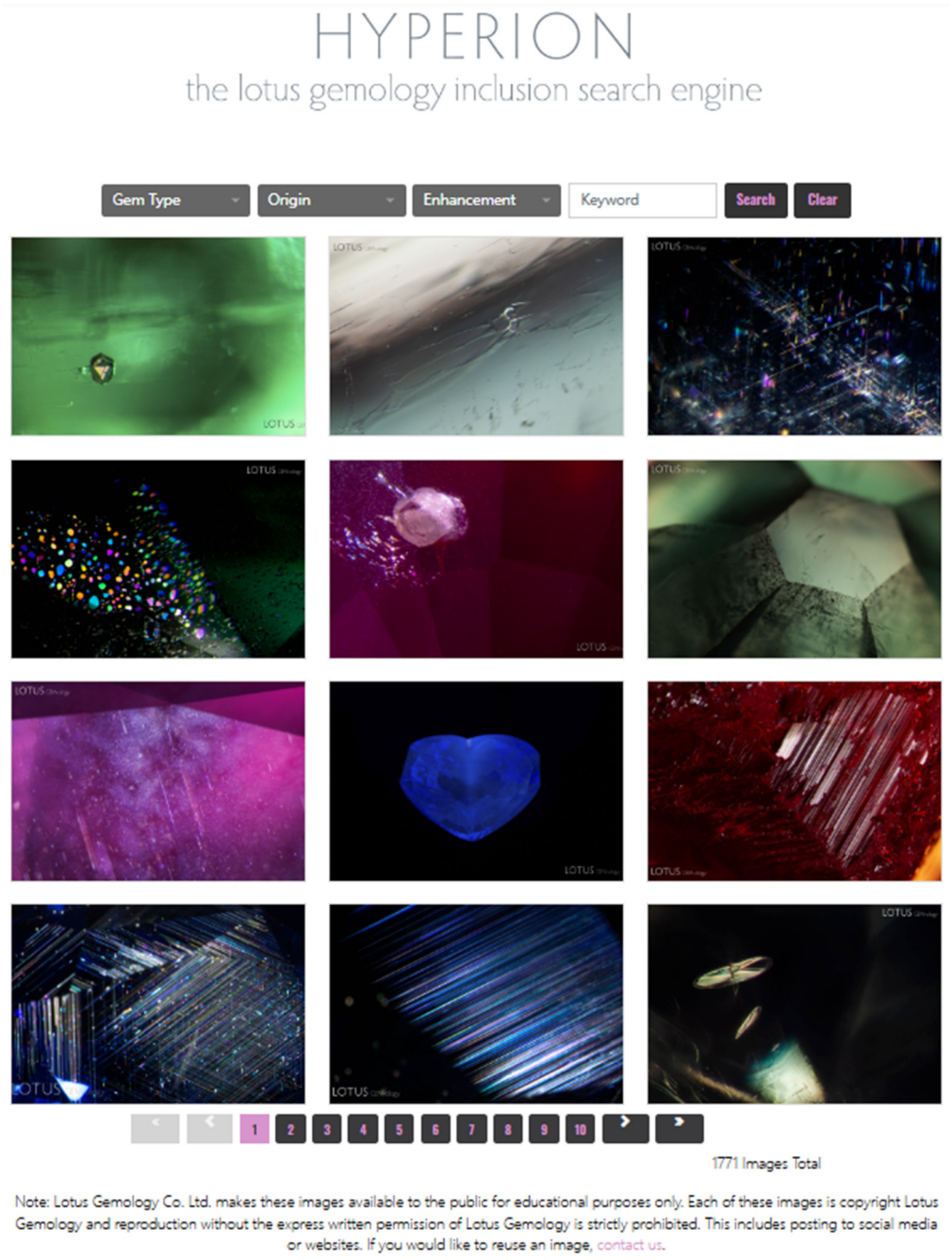



**Figure 5.** Hyperion Gemstone Inclusion Database, created by Lotus Gemology Laboratory.

These databases are essential for geological research, as they cover various types of data, including mineral composition, geochemical data, and spectral information. They are characterized by data diversity, standardized storage, and efficient retrieval functions, which provide a solid foundation for subsequent mineral research. However, these databases have insufficient gemstone sample collections and lack key gemstone information.

| Course Code | Sample ID | Gem Type | Color | Cut | Transparency | Luster | Size (mm) | Weight (ct) |
|---|---|---|---|---|---|---|---|---|
| FGA-B-14-23 | Ber.1101 | Beryl | Light green | Oval facet | Transparent | Glass luster | 5.88×4.06×3.33 | 1.22 |
| | Image |  | | | Dichroism | Weak, yellowish green/ blueish-green | | |
| | | | | | UV Fluorescence | LW: Inert, SW: Inert | | |
| | | | | | Chelsea Color Filter (CCF) | Green | | |
| | | | | | Polariscope | "Dark cross" uniaxial interference figure | | |
| | Inclusion |  | | | Feature Description | Two-phase gas-liquid parallel tubes with rain-like appearance (rain-like inclusions), irregular, jagged multiphase inclusions. | | |
| | RI | 1.565−1.573,DR=0.008 | | | | | | |
| | Spectroscope | A broad absorption band can be seen in the orange-red region, and an absorption line can be seen in the violet region. | | | | | | |
| | Note | Origin: Nigeria | | | | | | |

**Figure 6.** Gemstone information from the gemstone database of the School of Gemmology at China University of Geosciences (Beijing), using the beryl sample with ID Ber. 1101 as an example.

## 5. Discussion

### *5.1. Limitations of Existing Gemstone Databases*

Gemstones are highly prized for their rarity and aesthetic appeal, serving both as decorative items and as investment assets. Unlike natural mineral samples, gemstone samples include not only natural specimens, but also synthetic gemstones and those that have been enhanced or treated, including exposure to electromagnetic radiation, heating, or injection of oils or other substances. All of these are sold on the market as gemstones, and their values change significantly. Accurately determining the value of precious gemstones is critical to gemology because it is essential to maintaining consumer trust in the jewelry industry.

Some university and museum gemstone databases are primarily intended for education, and popular science. Only basic information such as specimen names, size, weight, appearance descriptions, and photographs is contained in these databases, and only simple search and management functions are offered. The core information critical to research, such as infrared spectroscopy, Raman spectroscopy, and ultraviolet spectroscopy analysis data, is lacking (Figure 6). These databases make very limited use of gemstone specimens, which is unfortunate because gemstones are a valuable resource. Well-known commercial jewelry testing institutions all have their own gemology databases, such as the Gemological Institute of America (GIA), the Gübelin Laboratory, the SSEF Laboratory, and the GRS Laboratory. However, commercial confidentiality prevents public access to these databases. While some gemstone data is included in various mineral databases, such as the International Mineralogical Association Database (IMA Database), the Mineral Database Mindat.org, and the Mineral Spectroscopy Database RRUFF Project, as well as China's

National Rock, Mineral, and Fossil Specimen Resource Sharing Platform, these databases primarily focus on extensive mineral information and specimen records. Nonetheless, these gemstone data are deficient in content that is of interest to gemological research institutes. Such content includes gemstone inclusion information, gemstone provenance, whether they have undergone treatment, quality grades, and information on synthetic gemstones.

Gemstone inclusions reveal crucial gemological features, such as internal and external characteristics of the gemstone. Often, microscopic magnification is required for observation. This information can be used to identify gemstone varieties, distinguish between natural and synthetic gemstones, determine the synthesis method of synthetic gemstones, and differentiate between treated and untreated gemstones. Perhaps most importantly, it can help determine the gemstone's origin. In the jewelry trade, the origin of high-value gemstones is a critical piece of information that significantly influences their value. Gemologists must be able to draw conclusions about origin based on scientific evidence, such as inclusions, fluorescence properties, spectral characteristics, and chemical composition distribution. Therefore, students require a large number of gemstone samples with reliable origin information during the stage of accumulating knowledge through teaching and research. By comparing the gemological characteristics of gemstones of the same variety from different deposits and geographical locations, students can learn to distinguish origins based on scientific evidence.

The development of gemstone synthesis technology is a vital component of modern gemology and can be traced back to the early 20th century. Technological advancements have led to the rapid proliferation of various gemstone synthesis methods. Replicating the color and appearance of natural gemstones through synthetic means has become a major research focus, significantly impacting the jewelry market. However, as synthesis techniques improve, the characteristics of synthetic gemstones change as well.

Additionally, gemstone treatment techniques continue to evolve, with new methods being developed to enhance the appearance and durability of gemstones. These techniques range from traditional methods like dying, filling, and heat treatment to newer diffusion techniques. Gemstones treated with new techniques (e.g., beryllium diffusion sapphires) often trigger alerts in gemological laboratories due to their subtle distinguishing characteristics. Researchers and gem labs are the first to capture this data.

Gemstone databases should promptly incorporate the latest data shared by researchers and laboratories to ensure timely updates to datasets. A more comprehensive identification system can be established by this database through the integration of big data technology, enhancing the accuracy, reliability, and timeliness of identification results.

Current issues with gemstone-related databases include incomplete data, inconsistent formatting standards, limited search functionality, a lack of uniform data standards across institutions, delayed data updates, and insufficient openness and sharing of data. These issues hinder the extraction of information from gemstone samples and limit the practical value of gemstone databases. The development of machine learning algorithms is contingent on access to high-quality mineral datasets, a crucial element in the current age of big data. Therefore, developing next-generation gemstone databases and fostering the synergistic development of deep learning and gemstone datasets is paramount.

### *5.2. Building a Novel Gemstone Database*

The methods used to identify gemstones traditionally depend on the experience of the person. These methods are subjective and have limited data for comparison. Advances in gemstone treatment techniques have made detecting such treatments increasingly difficult. This is true even for the most experienced human identifiers. Consequently, visual inspection alone is generally considered inadequate for reliable and repeatable treatment detection

and source identification in the modern setting. These challenges and industry demands have led advanced gemological laboratories to incorporate a suite of analytical instruments into their routine workflows. These instruments include ultraviolet-visible-near-infrared (UV), Fourier transform infrared (FTIR), energy-dispersive X-ray fluorescence (XRF), and laser ablation inductively coupled plasma mass spectrometry (ICP-MS) spectrometers. A vast amount of gem-related data has been generated by the use of these advanced instruments. Consequently, the construction of a new, comprehensive gemstone database while developing methods that effectively utilize this data and maximize accuracy and reliability is of significant practical importance.

The new gemstone database should be a data information system developed based on computer technology and geographic information systems (GIS) that adheres to the FAIR (Findability, Accessibility, Interoperability, Reusability) guiding principles [51]. It should encompass four major data structures: fundamental gemological data (e.g., physicochemical properties, optical characteristics, and provenance information); image data (e.g., photographs of samples, micrographs, and backscattered electron images); spectral and compositional data; and bibliographic data. The system should support standardized data processing and convenient search functionality.

In terms of application prospects, the new gemstone database can be expanded to include innovative features that will bring revolutionary changes to the gemstone industry. The integration of artificial intelligence technology will significantly increase the database's intelligence level. Machine learning algorithms can automatically identify gemstone characteristics, and deep learning models can learn complex identification rules from large sample sets. Data mining techniques can reveal hidden correlations between gemstone attributes, providing a scientific basis for origin tracing, quality assessment, and other applications. These improvements provide more efficient tools for jewelry testing and research institutions and enable mining companies to optimize exploration decisions, yielding significant cost savings. It allows human experts to focus on higher-value research and development. The establishment of a trustworthy data traceability mechanism is facilitated by blockchain technology, ensuring the authenticity and immutability of gemstone information. This enhances consumer confidence and promotes standardized market development. In the field of education, interactive learning systems based on gemstone databases can significantly improve teaching effectiveness.

We hope this new gemstone database meets the growing demand among researchers for internationally recognized, authoritative data, and fully supports and accelerates scientific discoveries.

### *5.3. Establish Appropriate Data Standards and Integrate Heterogeneous Data*

Traditional gemstone databases exhibit significant inadequacies in their data collections. These data sources of databases primarily rely on laboratory testing and limited market sampling, resulting in narrow coverage and slow updates. Globally, there are no widely recognized standards for collecting and exchanging gemstone data, and the mechanisms for sharing data are not adequate. This hinders interoperability between databases across institutions, creating severe data silos. Moreover, these data, which comprises textual descriptions, cartographic representations, tabular data, and photographic documentation, are present in a variety of formats. Data formats from the same experimental project may vary across different laboratories. In terms of data processing, traditional databases lack effective data mining and analysis tools, hindering the discovery of potential patterns and correlations within vast datasets. Traditional databases also exhibit weaknesses in user interaction experience and visualization, which diminishes their operational efficiency and educational value. Therefore, enhancing data acquisition efficiency and the intelligent

processing capabilities of gemstone databases are critical issues requiring urgent attention in current research.

Some gemstone test results depend solely on subtle variations in their physical, spectral, and chemical properties. Since gems are seen as long-term investments, they periodically re-enter the market, and the value of the same gem is frequently re-evaluated [12]. For this reason, it is critical to ensure the accuracy and reliability of gemstone identification and grading. A gemstone's value can be significantly impacted by discrepancies in determining its origin or treatment. This can, in turn, compromise the credibility of the entire asset class. Therefore, it is very important to follow strict laboratory rules and create strong standards to reduce personal opinions.

In gemological research, data encompasses the following types: (1) Fundamental gemological data, including physical and chemical properties, optical characteristics, and provenance information; (2) Image data, including specimen photographs, micrographs, and backscattered electron (BSE) images; (3) Laboratory test data, including results from spectroscopic analysis and compositional determination; and (4) Analogous data from literature references.

In order to integrate these heterogeneous data into a unified platform, the relationships between the data must first be clarified to guide database design. For instance, the origin of sapphires, which is a critical factor in their valuation, is not simply determined by the country of production. Rather, it is closely tied to specific geological structures and hydrothermal activity. The formation of sapphires can be primarily categorized into two types: metamorphic and magmatic [52,53]. Gem-quality sapphires are typically found in metamorphic rocks, such as those in Sri Lanka, Kashmir, and Myanmar. In contrast, sapphires of relatively lower market value originate from magmatic rocks and exhibit less desirable coloration, including those sourced from Thailand, Australia, and Shandong Province in China. The spatial and temporal correlation between tectonic settings and gem occurrences, as well as the relationship between gem quality and these settings, are critical considerations when constructing new gemstone databases.

Secondly, it is necessary to establish appropriate and unified data standards and technical terminology. This includes developing a scientific gemstone classification system, standardizing the identification process, and implementing a standardized data collection process. Currently, despite the existence of some classification systems, disputes and ambiguities persist in practical applications. The standardization of identification methods and data collection must be addressed by establishing an internationally recognized standardized identification system. These measures include the following: creating a set of standard evaluation guidelines and trial operating protocols; facilitating international appraisal training and knowledge-sharing; and encouraging the establishment of mutual recognition agreements among relevant nations. These steps are necessary for achieving global standardization in gemstone appraisal. This will foster the healthy development of the international gemstone trade.

Finally, data management and storage must adhere to the FAIR principles: findability, accessibility, interoperability, and reusability [51]. The database's architecture and functional design must consider aspects such as data storage security, query speed, and user-friendly interface. Integrating big data technologies enables real-time data updates and dynamic visualization, facilitating cross-domain and cross-level queries and analyses for users.

We propose constructing a novel gemstone database whose core objective is to overcome the limitations of traditional databases, such as a lack of standardised, machine learning-compatible data and insufficient feature integration, while also supporting the training, validation and iteration of machine learning models (e.g., DenseNet for grading

and Convolutional Neural Networks for synthetic detection). The new gemstone database's core framework comprises three interconnected layers: data acquisition, data processing, and application services. It features clear data specifications and a machine learning-oriented design. The specific functions and implementation methods for each layer are detailed in Table 2. The database is founded on the data acquisition layer, which integrates real-time data from various terminal devices and data sources. The middle layer is the data processing layer, which contains modules such as data cleansing, feature extraction, and pattern recognition. The top layer is the application service layer, which provides diverse functional interfaces. The database should use a hybrid storage solution that combines the advantages of relational and non-relational databases to meet the storage needs of different data types.

**Table 2.** Core framework of the novel gemstone database.

| Framework Layer | Core Components | ML-Oriented Design Specifications |
| --- | --- | --- |
| Data Acquisition Layer | Multi-source data collection modules (laboratory, industry, public) | (1) Image data: Standardized resolution (224 × 224 pixels for CNN input), unified lighting (6500 K white light) to avoid ML model bias from illumination variations.<br>(2) Spectral data: Normalized to 0–1 intensity range, resampled to 1 cm−1 intervals (for Transformer/LSTM sequence input), with background noise removed via Savitzky–Golay filtering.<br>(3) Label data: Annotated with ML task-specific tags (e.g., "clarity grade: VVS1" for diamond classification, "treatment type: beryllium diffusion" for detection), ensuring inter-annotator consistency. |
| Data processing layer | Automated feature extraction & fusion modules | (1) Image features: Pre-extracted via CNN or DenseNet (e.g., MobileNetV2) to output 128-dimensional feature vectors. For example, gemstone classification or grading, as well as determining the age and origin of archaeo-gemological materials through cutting and polishing techniques.<br>(2) Spectral features: Key parameters (peak position, intensity, half-width) extracted via Python scipy.signal, stored as structured arrays compatible with PyTorch/TensorFlow tensors.<br>(3) Multi-modal fusion: A dedicated module to align image, spectral, and geochemical features for multi-modal ML models. This module can be applied to trace the provenance (geological origin) of gemstones. For example, it can distinguish the origin of valuable gemstones such as rubies, sapphires, and emeralds. |
| Application service layer | Application program interfaces (API) & iterative feedback modules | (1) Data call API: RESTful interfaces supporting batch export of ML datasets (e.g., "export 5000 labeled diamond images and features"), with output formats (e.g., NumPy ''.npy", TensorFlow ''.tfrecord") customizable by model framework.<br>(2) Feedback loop: Real-time interface for ML models to return prediction results, which are linked to original data for expert review and retraining data update. |

This framework addresses the incompatibility between traditional gemstone databases and machine learning models. For instance, standardized image resolutions and pre-extracted features spare machine learning researchers from excessive data preprocessing time, while the multimodal fusion module lays the groundwork for training high-precision provenance (geographical origin) tracing models, which is a critical need in contemporary gemological research. Current machine learning-based gemstone identification models, such as GEMTELLIGENCE, primarily rely on rule-based algorithms that heavily depend on predefined physical–optical parameters or spectral composition features. In contrast,

the novel gemstone database not only supports machine learning but also integrates multi-source data through deep learning and blockchain technology for adaptive modeling. This approach demonstrates significant advantages in handling complex gemstone identification and revealing undisclosed treatment techniques.

## 6. Conclusions

Traditional gemstone databases have become inadequate in addressing the interdisciplinary research needs of gemology (e.g., provenance tracing, color genesis analysis) and the digital transformation of the gem industry, primarily due to three critical limitations: the absence of standardized data schemas, the lack of core datasets (e.g., synthetic/treatment records, inclusion characteristics), and poor data interoperability. Although existing mineral databases (e.g., Mindat.org, RRUFF Project) and proprietary databases developed by commercial entities have accumulated substantial mineralogical data, the former lack gemology-specific parameters (e.g., heat treatment signatures, market-oriented quality grading criteria), while the latter remain inaccessible due to commercial confidentiality. This "data silo" phenomenon significantly hinders the in-depth application of advanced technologies such as machine learning (ML) and artificial intelligence (AI) in gemstone identification and characterization.

An innovative gemstone database built on standardization (FAIR principles), integration (blockchain), and dynamism (real-time updates) provides a comprehensive solution for traditional defects. First, the gemstone database adheres to FAIR (Findability, Accessibility, Interoperability, Reusability) data principles, establishing a unified data schema encompassing gemstone basic properties, microstructural images (e.g., optical microscopy, backscattered electron microscopy), spectral datasets, and geochemical data. The new gemstone database's core framework comprises three interconnected layers: data acquisition, data processing, and application services. It features clear data specifications and will address the challenge of integrating heterogeneous data (e.g., unstructured text, tabular data, and spectral graphs). Second, blockchain technology should be incorporated to ensure data provenance and immutability, mitigating the risk of data tampering in critical applications such as gemstone authentication. Third, a real-time data update mechanism is required to incorporate emerging data to address challenges posed by evolving synthetic and treatment technologies. Current machine learning-based gemstone identification models primarily rely on rule-based algorithms that heavily depend on predefined physical-optical parameters or spectral composition features. In contrast, the novel gemstone database not only supports machine learning but also integrates multi-source data through deep learning and blockchain technology for adaptive modeling. This approach demonstrates significant advantages in handling complex gemstone identification and revealing undisclosed treatment techniques.

In the future, with big data and artificial intelligence technologies, the new gemstone database will provide reliable data support for gemological research. This will lead to intelligent upgrades in jewelry appraisal (reducing reliance on high-cost instruments), precision in mining exploration (optimizing resource decisions), and market standardization (enhancing consumer confidence through trustworthy data). Simultaneously, interactive learning systems will improve educational outreach. These will ultimately revolutionize the gemstone industry from an "experience-driven" to a "data-driven" model and will propel dual innovation in both academic research and practical applications.

**Author Contributions:** Conceptualization, Y.Z. and G.S.; methodology, G.S.; software, Y.Z.; validation, Y.Z. and G.S.; formal analysis, Y.Z.; investigation, Y.Z.; resources, G.S.; data curation, Y.Z.; writing—original draft preparation, Y.Z.; writing—review and editing, Y.Z. and G.S.; visualization,

Y.Z.; supervision, G.S.; project administration, G.S.; funding acquisition, Y.Z. and G.S. All authors have read and agreed to the published version of the manuscript.

**Funding:** This investigation was supported by 2025 Teaching Reform-Research and Application of Teaching Laboratory and Experimental Technology at China University of Geosciences (Beijing) (No. 640125002, Project SYJS202504).

**Data Availability Statement:** The data presented in this study are available within the article.

**Acknowledgments:** The authors thank Mingyue He and Mei Yang (National Infrastructure of Mineral, Rock and Fossil for Science and Technology) for the technical support. The authors thank the Museum of China University of Geosciences (Beijing) for its support.

**Conflicts of Interest:** The authors declare no conflicts of interest.

## Abbreviations

The following abbreviations are used in this manuscript:

| | |
|---|---|
| GIA | Gemological Institute of America |
| SSEF | Swiss Gemmological Institute |
| GRS | GemResearch Swisslab |
| Gübelin | Gübelin Gemmological Laboratory |

## References

1. Zhao, X.; Wang, Q.; Pham, H. Big data modeling and applications. *Ann. Oper. Res.* **2025**, *348*, 1. [CrossRef]
2. Liu, B.; Zhai, M. Prospects for Construction New Metamorphic Rock Database in Big Data Epoch. *J. Earth Sci.* **2025**, *36*, 450–459. [CrossRef]
3. Guo, H. Scientific Big Data—A Footstone of National Strategy for Big Data. *Bull. Chin. Acad. Sci.* **2018**, *33*, 768–773. [CrossRef]
4. Chang, K.; Lee, S.; Park, J.; Chung, H. Feasibility for non-destructive discrimination of natural and beryllium-diffused sapphires using Raman spectroscopy. *Talanta* **2016**, *149*, 335–340. [CrossRef]
5. Zhang, Y.; Shi, G.; Hao, Y. Nondestructive and quick differentiation between loupe-clean low-alkali natural and hydrothermal synthetic aquamarines: Constraints of infrared spectroscopy. *J. Solid State Chem.* **2024**, *340*, 125044. [CrossRef]
6. Zhang, Y.; Shi, G.; Xie, Z. Spectral Characteristics of Nitrogen-Doped CVD Synthetic Diamonds and the Origin of Surface Blue Fluorescence. *Crystals* **2024**, *14*, 804. [CrossRef]
7. Zhang, Y.; Shi, G.; Wen, J. Chromite and Its Thin Kosmochlor and Cr-Omphacite Cortex in Amphibolite from the Myanmar Jadeite Deposits. *Crystals* **2025**, *15*, 79. [CrossRef]
8. Zhang, Y.; Shi, G.; Wen, J. Manganese-Rich Chromite in Myanmar Jadeite Jade: A Critical Source of Chromium and Manganese and Its Role in Coloration. *Crystals* **2025**, *15*, 704. [CrossRef]
9. Wang, C.; Hazen, R.M.; Cheng, Q.; Stephenson, M.H.; Zhou, C.; Fox, P.; Shen, S.-z.; Oberhänsli, R.; Hou, Z.; Ma, X.; et al. The Deep-Time Digital Earth program: Data-driven discovery in geosciences. *Natl. Sci. Rev.* **2021**, *8*, 156–166. [CrossRef] [PubMed]
10. Alonso-Perez, R.; Day, J.M.D.; Pearson, D.G.; Luo, Y.; Palacios, M.A.; Sudhakar, R.; Palke, A. Exploring emerald global geochemical provenance through fingerprinting and machine learning methods. *Artif. Intell. Geosci.* **2024**, *5*, 100085. [CrossRef]
11. Zhang, Y.; Shi, G. Origin of Blue-Water Jadeite Jades from Myanmar and Guatemala: Differentiation by Non-Destructive Spectroscopic Techniques. *Crystals* **2022**, *12*, 1148. [CrossRef]
12. Bendinelli, T.; Biggio, L.; Nyfeler, D.; Ghosh, A.; Tollan, P.; Kirschmann, M.A.; Fink, O. GEMTELLIGENCE: Accelerating gemstone classification with deep learning. *Commun. Eng.* **2024**, *3*, 110. [CrossRef]
13. Pena, F.B.; Crabi, D.; Izidoro, S.C.; Rodrigues, É.O.; Bernardes, G. Machine learning applied to emerald gemstone grading: Framework proposal and creation of a public dataset. *Pattern Anal. Appl.* **2022**, *25*, 241–251. [CrossRef]
14. Neva-Rodríguez, D.; Ochoa-Gutierrez, L.H. Estimation of emerald mineralization probability using machine learning algorithms. *DYNA* **2025**, *92*, 19–27. [CrossRef]
15. Culka, A.; Jehlička, J. Identification of gemstones using portable sequentially shifted excitation Raman spectrometer and RRUFF online database: A proof of concept study. *Eur. Phys. J. Plus* **2019**, *134*, 130. [CrossRef]
16. Breeding, C.M.; Shigley, J.E. The "Type" Classification System of Diamonds and Its Importance in Gemology. *Gems Gemol.* **2009**, *45*, 96–111. [CrossRef]
17. Fisher, D.; Evans, D.J.F.; Glover, C.; Kelly, C.J.; Sheehy, M.J.; Summerton, G.C. The vacancy as a probe of the strain in type IIa diamonds. *Diam. Relat. Mater.* **2006**, *15*, 1636–1642. [CrossRef]

18. Bogert, C.H.V.D.; Smith, C.P.; Hainschwang, T.; McClure, S.F. Gray-to-Blue-to-Violet Hydrogen-Rich Diamonds from The Argyle Mine, Australia. *Gems Gemol.* **2009**, *45*, 20–37. [CrossRef]
19. Wang, W.; D'Haenens-Johansson, U.F.S.; Johnson, P.; Moe, K.S.; Emerson, E. CVD Synthetic Diamonds from Gemesis Corp. *Gems Gemol.* **2012**, *48*, 80–97. [CrossRef]
20. Wang, W.Y.; Moses, T.; Linares, R.C.; Shigley, J.E.; Hall, M.; Butler, J.E. Gem-quality synthetic diamonds grown by a chemical vapor deposition (CVD) method. *Gems Gemol.* **2003**, *39*, 268–283. [CrossRef]
21. Martineau, P.M.; Lawson, S.C.; Taylor, A.J.; Quinn, S.J.; Evans, D.J.F.; Crowder, M.J. Identification of synthetic diamond grown using chemical vapor deposition (CVD). *Gems Gemol.* **2004**, *40*, 2–25. [CrossRef]
22. Meng, Y.F.; Yan, C.S.; Lai, J.; Krasnicki, S.; Shu, H.; Yu, T.; Liang, Q.; Mao, H.K.; Hemley, R.J. Enhanced optical properties of chemical vapor deposited single crystal diamond by low-pressure/high-temperature annealing. *Proc. Natl. Acad. Sci. USA* **2008**, *105*, 17620–17625. [CrossRef] [PubMed]
23. Saeseaw, S.; Pardieu, V.; Sangsawong, S. Three-Phase Inclusions in Emerald and Their Impact on Origin Determination. *Gems Gemol.* **2014**, *50*, 114–132. [CrossRef]
24. Karampelas, S.; Al-Shaybani, B.; Mohamed, F.; Sangsawong, S.; Al-Alawi, A. Emeralds from the Most Important Occurrences: Chemical and Spectroscopic Data. *Minerals* **2019**, *9*, 561. [CrossRef]
25. Saeseaw, S.; Renfro, N.D.; Palke, A.C.; Ziyin, S.; McClure, S.F. Geographic origin determination of emerald. *Gems Gemol.* **2019**, *55*, 614–646. [CrossRef]
26. Nicol, C.-A.; Marshall, D.; Einfalt, H.C.; Thorkelson, D. Pressure-Temperature-Fluid Constraints for the Formation of the Halo-Shakiso Emerald Deposit, Southern Ethiopia: Fluid Inclusion and Stable Isotope Studies. *Can. Mineral.* **2022**, *60*, 29–48. [CrossRef]
27. Pardieu, V.; Sangsawong, S.; Luetrakulprawat, S.; Cornuz, L.; Raynaud, V. Update on Emeralds from the Mananjary-Irondro Area, Madagascar. *J. Gemmol.* **2020**, *37*, 416–425. [CrossRef]
28. Nikopoulou, M.; Karampelas, S.; Hennebois, U.; Gruss, P.; Gaillou, E.; Fritsch, E.; Herreweghe, A.; Papadopoulou, L.; Melfos, V.; Kantiranis, N.; et al. Microscopic, Spectroscopic and Chemical Analysis of Emeralds from Habachtal, Austria. *Minerals* **2024**, *15*, 22. [CrossRef]
29. Franz, G.; Vyshnevskyi, O.; Taran, M.; Khomenko, V.; Wiedenbeck, M.; Schiperski, F.; Nissen, J. A new emerald occurrence from Kruta Balka, Western Peri-Azovian region, Ukraine: Implications for understanding the crystal chemistry of emerald. *Am. Mineral.* **2020**, *105*, 162–181. [CrossRef]
30. Krzemnicki, M.S.; Wang, H.A.O.; Büche, S. A New Type of Emerald from Afghanistan's Panjshir Valley. *J. Gemmol.* **2021**, *37*, 474–495. [CrossRef]
31. Khaleal, F.M.; Saleh, G.M.; Lasheen, E.S.R.; Lentz, D.R. Occurrences and genesis of emerald and other beryls mineralization in Egypt: A review. *Phys. Chem. Earth Parts A/B/C* **2022**, *128*, 103266. [CrossRef]
32. Hanser, C.S.; Stephan, T.; Gul, B.; Häger, T.; Botcharnikov, R. Comparison of Emeralds from the Chitral District, Pakistan, with other Pakistani and Afghan Emeralds. *J. Gemmol.* **2023**, *38*, 582–599. [CrossRef]
33. Satapathy, J.S.; Singh, S.; Sahoo, P.R. Mineralogical and Geochemical Characteristics of Emeralds from the Bahutiya and Gurabanda Deposits of Jharkhand, India, and Comparison with Other World Emerald Occurrences. *Acta Geol. Sin. Engl. Ed.* **2025**, *99*, 553–567. [CrossRef]
34. Zenetos, M.C. Update on Emerald Mining at Campos Verdes, Goiás, Brazil. *J. Gemmol.* **2022**, *38*, 312–313. [CrossRef]
35. Monarumit, N.; Lhuaamporn, T.; Sakkaravej, S.; Wathanakul, P.; Wongkokua, W. The color center of beryllium-treated yellow sapphires. *J. Phys. Commun.* **2020**, *4*, 105018. [CrossRef]
36. Emmett, J.L.; Scarratt, K.; McClure, S.F.; Moses, T.; Douthit, T.R.; Hughes, R.; Novak, S.; Shigley, J.E.; Wang, W.; Bordelon, O. Beryllium Diffusion of Ruby and Sapphire. *Gems Gemol.* **2003**, *39*, 84–135. [CrossRef]
37. Butini, E.; Butini, F.; Angle, M.; Cerino, P.; De Angelis, A.; Tomei, N.; Altamura, F. Archaeometric and gemmological analyses of a Roman imperial gold-and-sapphire jewel from Colonna (Rome, Italy). *Measurement* **2018**, *128*, 160–169. [CrossRef]
38. Saeseaw, S.; Sangsawong, S.; Vertriest, W.; Atikarnsakul, U.; Liliane, V.; Raynaud-Flattot Khowpong, C.; Weeramonkhonlert, V. A Study of Sapphire from Chanthaburi, Thailand and its Gemological Characteristics. *Gems Gemol.* **2017**, *53*, 1–42. Available online: https://www.gia.edu/gia-news-research/sapphire-chanthaburi-thailand-gemological-characteristics (accessed on 27 October 2025).
39. Saminpanya, S.; Manning, D.A.C.; Droop, G.T.R. Trace elements in Thai gem corundums. *J. Gemmol.* **2003**, *28*, 399–415. [CrossRef]
40. Freire, W.M.; Amaral, A.M.M.M.; Costa, Y.M.G. Gemstone classification using ConvNet with transfer learning and fine-tuning. In Proceedings of the 2022 29th International Conference on Systems, Signals and Image Processing (IWSSIP), Sofia, Bulgaria, 1–3 June 2022; pp. 1–4.
41. Chakraborty, B.; Mukherjee, R.; Das, S. Gemstone Classification Using Deep Convolutional Neural Network. *J. Inst. Eng. India Ser. B* **2024**, *105*, 773–785. [CrossRef]

42. Huang, G.; Liu, Z.; Pleiss, G.; Maaten Lvd Weinberger, K.Q. Convolutional Networks with Dense Connectivity. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *44*, 8704–8716. [CrossRef]
43. Chow, B.; Reyes-Aldasoro, C. Automatic Gemstone Classification Using Computer Vision. *Minerals* **2021**, *12*, 60. [CrossRef]
44. Ostreika, A.; Pivoras, M.; Misevičius, A.; Skersys, T.; Paulauskas, L. Classification of Objects by Shape Applied to Amber Gemstone Classification. *Appl. Sci.* **2021**, *11*, 1024. [CrossRef]
45. Sinkevicius, S.; Lipnickas, A.; Rimkus, K. Automatic amber gemstones identification by color and shape visual properties. *Eng. Appl. Artif. Intell.* **2015**, *37*, 258–267. [CrossRef]
46. Wang, D.; Bischof, L.; Lagerstrom, R.; Hilsenstein, V.; Hornabrook, A.; Hornabrook, G. Automated Opal Grading by Imaging and Statistical Learning. *IEEE Trans. Syst. Man Cybern. Syst.* **2016**, *46*, 185–201. [CrossRef]
47. Khalilian, P.; Rezaei, F.; Darkhal, N.; Karimi, P.; Safi, A.; Palleschi, V.; Melikechi, N.; Tavassoli, S.H. Jewelry rock discrimination as interpretable data using laser-induced breakdown spectroscopy and a convolutional LSTM deep learning algorithm. *Sci. Rep.* **2024**, *14*, 5169. [CrossRef]
48. Hyperion Gemstone Inclusion Database, Created by Lotus Gemology Laboratory. Available online:https://lotusgemology.com/index.php/en/resources/hyperion-inclusion-repository (accessed on 12 September 2025).
49. Ma, X.; Ralph, J.; Zhang, J.; Que, X.; Prabhu, A.; Morrison, S.M.; Hazen, R.M.; Wyborn, L.; Lehnert, K. OpenMindat: Open and FAIR mineralogy data from the Mindat database. *Geosci. Data J.* **2023**, *11*, 94–104. [CrossRef]
50. Lafuente, B.; Downs, R.T.; Yang, H.; Stone, N. The power of databases: The RRUFF project. In *Highlights in Mineralogical Crystallography*; Armbruster, T., Danisi, R.M., Eds.; De Gruyter: Berlin, Germany, 2015; pp. 1–30.
51. Wilkinson, M.D.; Dumontier, M.; Aalbersberg, I.J.; Appleton, G.; Axton, M.; Baak, A.; Blomberg, N.; Boiten, J.-W.; da Silva Santos, L.B.; Bourne, P.E.; et al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* **2016**, *3*, 160018. [CrossRef] [PubMed]
52. Belley, P.M.; Onodenalore, O.; Abdale, L.; Groat, L.A.; Fayek, M. Chemical discrimination of magmatic vs. metamorphic blue corundum: The problem of corundum formed in partial melt and its implications for sapphire exploration and deposit modeling. *Lithos* **2025**, *510–511*, 108125. [CrossRef]
53. Giuliani, G.; Groat, L.A. Geology of Corundum and Emerald Gem Deposits: A Review. *Gems Gemol.* **2019**, *55*, 464–489. [CrossRef]


**Disclaimer/Publisher's Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.