nanomaterials

MDPI

Article

Machine Learning Analysis of Raman Spectra of MoS_2

Yu Mao [1,2], Ningning Dong [1,2,*], Lei Wang [1,2], Xin Chen [1,2], Hongqiang Wang [1,2], Zixin Wang [1,2], Ivan M. Kislyakov [1,2] and Jun Wang [1,2,3,4,*]

1 Laboratory of Micro-Nano Optoelectronic Materials and Devices, Key Laboratory of Materials for High-Power Laser, Shanghai Institute of Optics and Fine Mechanics, Chinese Academy of Sciences, Shanghai 201800, China; yumao@siom.ac.cn (Y.M.); wanglei2016@siom.ac.cn (L.W.); XinChen@siom.ac.cn (X.C.); hqwang@siom.ac.cn (H.W.); zxwang@siom.ac.cn (Z.W.); iv.kis@mail.ru (I.M.K.)

2 Center of Materials Science and Optoelectronics Engineering, University of Chinese Academy of Sciences, Beijing 100049, China

3 State Key Laboratory of High Field Laser Physics, Shanghai Institute of Optics and Fine Mechanics, Chinese Academy of Sciences, Shanghai 201800, China

4 CAS Center for Excellence in Ultra-Intense Laser Science (CEULS), Shanghai 201800, China

* Correspondence: n.n.dong@siom.ac.cn (N.D.); jwang@siom.ac.cn (J.W.)

Received: 28 September 2020; Accepted: 6 November 2020; Published: 9 November 2020

Abstract: Defects introduced during the growth process greatly affect the device performance of two-dimensional (2D) materials. Here we demonstrate the applicability of employing machine-learning-based analysis to distinguish the monolayer continuous film and defect areas of molybdenum disulfide (MoS_2) using position-dependent information extracted from its Raman spectra. The random forest method can analyze multiple Raman features to identify samples, making up for the problem of not being able to effectively identify by using just one certain variable with high recognition accuracy. Even some dispersed nucleation site defects can be predicted, which would commonly be ignored under an optical microscope because of the lower optical contrast. The successful application for classification and analysis highlights the potential for implementing machine learning to tap the depth of classical methods in 2D materials research.

Keywords: 2D materials; machine learning; random forest algorithm; Raman spectrum

1. Introduction

Transition-metal dichalcogenides (TMDCs) are a class of layered materials analogous to graphene, which have aroused immense interest in the last decade as a promising platform for electronic and optoelectronic applications in the post-Moore era [1,2]. The adjacent layers in these materials are held together by weak van der Waals forces but strong covalent bonding forces inside the layer, making it possible to cleave or synthetize to the limit of a monolayer. Even though TMDCs have been studied for decades in their bulk form, the properties of monolayers and few-layers with ultrathin thickness differ dramatically from the macroscopic material characteristics, being the biggest reason for renewed interest in this material class. Compared to their bulk counterpart, monolayers of many TMDCs (such as MoS_2, WS_2, WSe_2 and $MoSe_2$) are especially exciting since they are direct band gap semiconductors, making them ideal candidates to replace silicon for device applications, such as light-emitting diodes, photodetectors and photodiodes [3]. The chemical vapor deposition (CVD) method provides a convenient and controllable way to grow high-quality and large area 2D materials at a reasonable cost, which has been earmarked as the process that will deliver scalable production [4,5]. Nowadays, continuous films of 2D materials compatible with current silicon-based microfabrication

processes are greatly needed for industrial electronic and optoelectronics applications [6–8]. However, the present uniformity of as-grown monolayer films is still inadequate, such as structural differences and poor controllable layer distribution, etc. It is a remarkable fact that the introduced cracks that appear in synthesis processes can adversely affect the device performance and directly increase the chip failure risk [9]. These factors greatly limit the further applications of these materials. As a consequence, it is essential to check the uniformity of the obtained 2D materials. It is necessary to locate and distinguish the monolayer continuous films and the random crack areas, as well as bilayer areas prone to be introduced in the growth process, in order to better understand and improve the growth process [10].

To identify these areas with thickness differences, the most typical methods are atomic force microscopy (AFM) [11,12], optical microscopy (OM) [13–15], differential reflectance spectra [16,17] and Raman spectra [18,19]. AFM is a versatile method used to measure the thickness of 2D materials; however, the materials are easily destroyed due to the sliding of the tip on the sample surface when working on contact mode, while the tapping mode takes a relatively long time to measure even a small area [20,21]. Time-domain terahertz spectroscopy is an emerging technique for imaging 2D materials, but the comparatively large spot size hinders the identification of laterally small flakes [22]. In such a situation, an accurate, versatile and nondestructive method is highly desirable not only in fundamental research but also for practical applications. Recently, machine learning approaches have attracted considerable attention for solving various problems in materials science and optical engineering. OM using red–green–blue (RGB)-based optical contrast combined with machine learning has shown emerging potential in identifying and determining the thicknesses of 2D materials. Lin et al. first used the support vector machine (SVM) method to learn the contrast information of optical images to determine the layer numbers of graphene and MoS_2 [23]. Later, clustering analysis [24] and the convolutional neural network (CNN) [25–27] also joined in this stage play, expanding the identification types and application scenarios of 2D materials. However, because the accuracy of using optical images to determine the layer number based on machine learning is not too high, the previous researchers mainly regarded it as an initial screening to reduce manual work [25]. The attempt to use photoluminescence (PL) imaging and computer vision techniques to analyze monolayer TMDCs also enlightened us to combine the research methods of intrinsic material properties with new technologies to obtain information about molecular structure and layer number simultaneously [28]. Nowadays, based on the relation between the Raman frequency shifts of the E^1_{2g} and A_{1g} peaks (TMDCs) and those of the 2D and G peaks (graphene), Raman spectroscopic mapping has been widely used to identify the thickness and confirm the uniformity of 2D materials [29–32]. As a high-resolution imaging technique, it does give us more information to study the nature of matter. However, it is difficult to identify these defects in the selected area and at the same time simply using one-side Raman frequency shifts. Hence, we consider solving this problem by introducing a machine learning approach to use more Raman features for the simultaneous identification.

To our knowledge, we are the first to present a recognition method to distinguish the monolayer continuous film and random defect areas of 2D semiconductors using the machine learning method with Raman signals. Compared to other unsupervised techniques, the supervised machine learning represented by random forest can not only reduce the computational expense and time but also achieves high accuracy. In the introduction process for the random forest algorithm, we use several Raman characteristics extracted from spatial mapping results as the input variables and the sample thickness type as the output variable for generating the decision trees. The successful application of a machine learning approach to the classification and analysis of the CVD-prepared MoS_2 highlights the potential of this method for 2D materials research.

2. Materials and Methods

2.1. Synthesis of Monolayer MoS_2 Continuous Film

The monolayer MoS_2 continuous film mentioned in this paper was synthesized via the CVD method similar to our previous work [33]. The film was grown using MoO_3 powders (99.97% Sigma Aldrich, St. Louis, MO, USA) as the molybdenum source and sulfur powders (99.98%, Sigma-Aldrich, St. Louis, MO, USA). First, MoO_3 and sulfur powders were loaded into two separate Al_2O_3 crucibles, which were located at the center and the upstream of a dual-temperature-zone tube furnace with a diameter of 100 mm. A piece of Si substrate with thermally grown 300-nm-thick SiO_2 was loaded at the downstream. Before the film growing, the quartz tube was evacuated to 4000 Pa at room temperature. Then, the temperature of the MoO_3 was increased to 670 °C and the temperature of the sulfur was increased to 190 °C with 50 sccm of argon gas, and maintained for 10 min. Finally, the furnace was naturally cooled down to room temperature.

2.2. Characterization and Measurements

We carried out PL, AFM and Raman measurements. The PL signals were collected by a confocal microscopy setup (LabRAM HR Evolution, Horiba Co., Kyoto, Japan) with a 532 nm continuous-wave (CW) laser of a frequency-doubled Nd:YAG laser. The height profiles were measured using AFM taken by an FM-Nanoview6800 (FSM-Precision Co., Suzhou, China) in tapping m ode. Raman spectra were obtained using the same confocal microscopy system equipped with a programmable scanning stage with a 532 nm CW laser (Changchun New Industries Optoelectronics Tech. Co., Changchun, China) as the excitation source. We chose the 100× objective lens (MPLFLN 100×, NA = 0.9) and set the laser power below 1 mW to avoid local heating and undesirable oxidation of the sample. The scanning step was set as 0.2 μm and the integration time was also carefully optimized to obtain an adequate spectrum resolution and a satisfactory signal-to-noise ratio, while maintaining acceptable data acquisition duration and avoiding drift.

3. Results and Discussion

As shown in Figure 1a, the MoS_2 monolayer continuous film has a relatively smooth surface. The thickness of the monolayer region is around 0.88 nm as confirmed by AFM, which corresponds to an interlayer S–Mo–S layer [19]. The heights of the undertint line and dark triangle areas were found to be crack and bilayer defects, respectively. The PL spectra in Figure 1b show two peaks at 670 nm and 620 nm corresponding to A (1.9 eV) and B (2.0 eV) direct excitonic transitions with the energy split from the valence-band spin–orbital coupling, respectively. The bilayer shows a decline in PL intensity compared with the monolayer [34]. As shown in Figure 1c, two Raman-active modes, E^1_{2g} and A_{1g}, exhibit significant differences, and the Raman frequency difference of the monolayer sample is ~17.9 cm^{-1}, while that of the bilayer sample is ~21.1 cm^{-1}. This is consistent with previous results for mechanical exfoliation (ME)-prepared and CVD-prepared samples, implying this difference is universal between samples obtained from different preparation methods [35–37]. The cracks nucleated at sulfur vacancies propagate along the energy-favored zigzag directions upon the relatively fast temperature-drop-induced thermal strain, which results in an orientation-specific fracture behavior [38]. In addition, the appearance of an E^1_{2g} mode proves that all the monolayer and bilayer samples are 2H-MoS_2 [39]. In the completely exposed internal area of the crack, the Raman signal of the Si substrate is mainly collected. The peak around 520 cm^{-1} is attributed to the Si mode. Similar to the case of multilayer graphene, the Raman signals from the Si substrate can be absorbed by the MoS_2 flakes, which makes the intensity of the Si mode monotonously decrease from the bare substrate to the monolayer and bilayer MoS_2 flakes [40,41]. In Figure 1c, the intensity of the Si mode is normalized to display the spectral information more intuitively like the previous work [42].

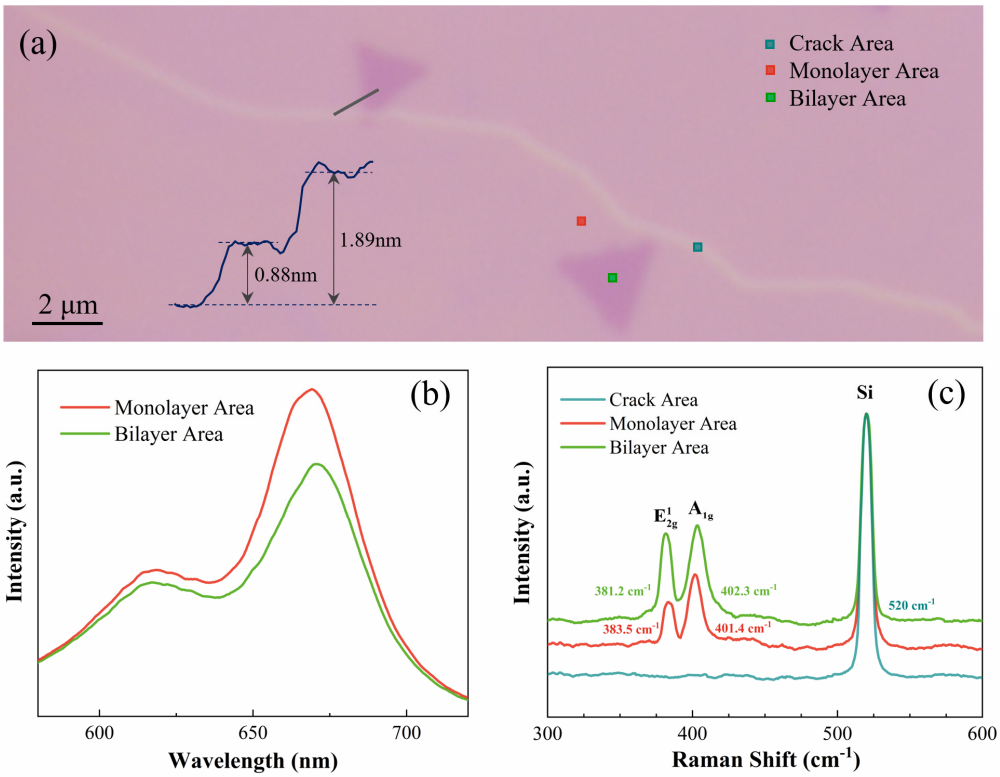

Figure 1. (**a**) Optical image of the MoS_2 sample. The inset shows the height profile, and the atomic force microscopy (AFM) profile is taken along the gray line drawn on the optical image. (**b**) Photoluminescence (PL) spectra of the monolayer and bilayer areas. (**c**) Raman spectra of the monolayer, crack and bilayer areas.

The pixels of the optical image in Figure 1a were reduced to be consistent with the collection points of the spectra by bicubic sharper process. Then, the reserved pixels could be clustered using the k-means algorithm [43], which can partition all pixels into three clusters with each cluster having a mean value, and pixels in one cluster are closest to the corresponding mean value among the cluster means [44]. Combined with the AFM heights, three regions with crack, monolayer and triangle bilayers can be identified as shown in Figure 2a. When using k-means clustering to classify different regions of the optical image, in addition to ensuring sufficient optical contrast, the clustering also depends on the AFM data to ensure effective classification. Nowadays, spectroscopy is the backbone of research in such diverse fields, ranging from physics to engineering, chemistry and biology [45]. Compared to relying on AFM to determine thickness, the more convenient way is to use the potentiality of classical spectroscopy research methods to obtain information about molecular structure, chemical composition and even layer number simultaneously.

In the research methods for 2D material properties, Raman spectroscopy is the most commonly used technique in many fields, since it allows the essential characteristics of matter that are invisible by standard OM and AFM to be viewed [46]. Ultralow-frequency Raman spectroscopy has been used to reliably determine layer numbers of TMDC flakes, but this technique requires expensive adapters and nonstandard equipment setup [47]. Therefore, it is of vital importance to look for a suitable technique using the standard Raman system. Generally, high-frequency Raman peaks of lattice vibrations (i.e., phonons) in TMDCs exhibit several prominent features, including frequency, intensity and full width at half maximum, which contain useful information in characterizing the physical and chemical properties of the materials. Figure 2b,c show the Raman spectral mapping of the E^1_{2g} and the A_{1g} peaks. From these two mappings, we can find the differences between the monolayer and bilayer areas. The frequency of the E^1_{2g} peak decreases, while that of the A_{1g} peak increases with increasing layer number. The observed blueshift of the A_{1g} peak originates from the constraint of

atom vibration by the interlayer van der Waals force in MoS_2, whereas the stacking-induced structure changes or long-range Coulomb interlayer interactions account for the redshift of the E^1_{2g} peak [18]. Therefore, as shown in Figure 2d, the Raman frequency difference between the E^1_{2g} and A_{1g} peaks can be used as a fingerprint feature to identify the monolayer and bilayer MoS_2 regions of the flakes [19]. However, the crack area cannot be identified straightforwardly by only using the frequency shifts of the E^1_{2g} or A_{1g} peaks.

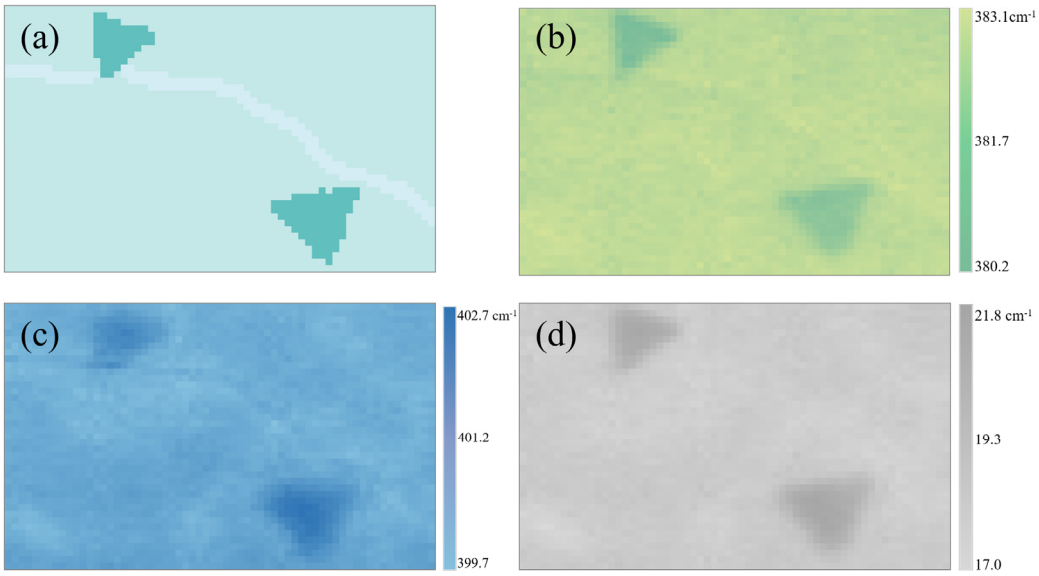

Figure 2. (**a**) K-means algorithm clustered image for the selected sample with monolayer (cyan), bilayer (dark cyan), and crack (light cyan) regions. Raman spectral mapping of the (**b**) Pos(E^1_{2g}) and the (**c**) Pos(A_{1g}). (**d**) Raman spectral mapping of the frequency difference between the Pos(E^1_{2g}) and the Pos(A_{1g}).

The great discrepancies in Raman intensity from crack to monolayer inspired us to consider more features extracted from Raman spectra to realize the simultaneous identification of these three regions. It is worth noting that when the detection position controlled by the scanning stage moves to the boundary of two regions, i.e., the junction of the crack and monolayer, it will inevitably collect signals both from the crack and monolayer MoS_2 at the same time, which makes it difficult to classify different areas based on Raman intensity by setting the threshold manually. We chose the average value of the E^1_{2g} peak intensity of a large monolayer MoS_2 area as the reference intensity, and all thresholds (a fixed percentage) were set based on this value. The Raman spectral intensity mappings of the E^1_{2g} peak are shown in Figure 3a–c and the thresholds are 68, 70 and 72%, respectively. It is not difficult to see that when the threshold changed slightly, the predicted information of the crack area and the monolayer area also changed. This means that an effective evaluation tool is lacked for judging the rationality of the artificially set thresholds. In addition, the monolayer sample generally has two Raman peaks, E^1_{2g} and A_{1g}, under the excitation of green or blue lasers with proper power. When using the same intensity threshold of the A_{1g} peak to make judgments, there are also some differences in the conclusions drawn compared to the E^1_{2g} peak, as shown in Figure 3d–f, which further increase the difficulty of manual processing. The rising machine learning approaches can help to successfully extract and analyze the multiple Raman characteristics among many samples to address this problem.

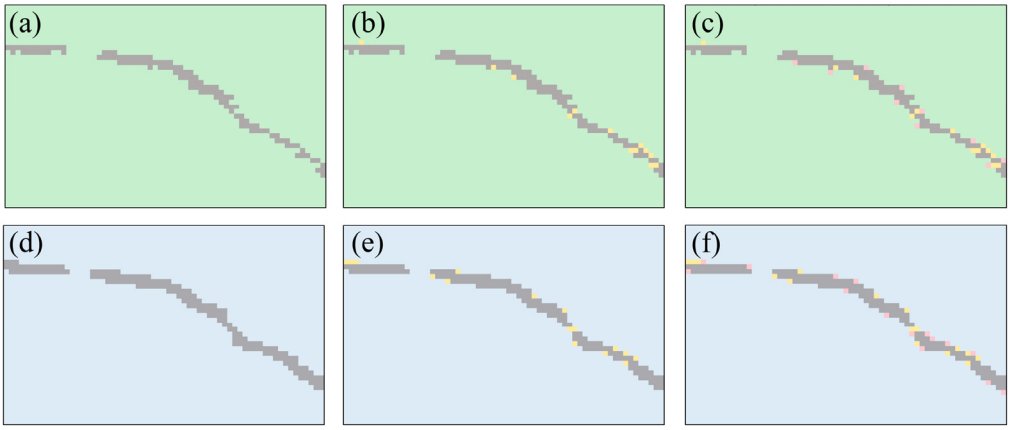

Figure 3. Raman spectral intensity mapping of the E^1_{2g} and the A_{1g} peaks. Gray areas and the newly added yellow and red areas indicate the situation when the E^1_{2g} and the A_{1g} peak intensity drops below (**a**,**d**) 68%, (**b**,**e**) 70% and (**c**,**f**) 72% of the monolayer signal, respectively.

The introduction of machine learning enables computers to tackle problems involving knowledge of the real world and make decisions that appear correct. Here we implement the random forest algorithm (Figure 4) to search for a hidden correlation that may exist between the sample types and the characteristic data obtained from the spatial Raman mapping. This method has been successfully applied on a PL spectra study [48]. Compared to other classification procedures, the random forest machine learning approach has the advantage of high classification accuracy [49]. Furthermore, it can determine variable importance and model complex interactions among predictor variables [50]. Here, we define z^i, including five types of spectral information (α, β, γ, δ and ε) as the input variables: α and β are the intensity and frequency of the E^1_{2g} peak, respectively; γ and δ are the intensity and frequency of the A_{1g} peak, respectively; ε is the Raman frequency difference between the two peaks previously mentioned. The sample types of crack, monolayer and bilayer are defined as output variables, which are acquired from the k-means algorithm results. The $Z^{Orig} = [z^1, z^2, z^3, \ldots, z^i, \ldots, z^k]$ are all training data for machine learning. Bootstrap sampling is used to expand a moderate number of data sets into a large volume of data sets required to improve the classification accuracy, which is a resampling technique used to estimate statistics on a population by sampling a dataset with replacement. Concretely speaking, we create n decision trees by generating new data sets Z^1, $Z^2, \ldots, Z^n$, using n data sets randomly extracted from the original Z^{Orig} with duplication permitted. For example, $Z^1 = [z^3, z^7, z^{16}, z^g, \ldots, z^i]$ and $Z^2 = [z^5, z^5, z^{15}, \ldots, z^k]$, where g and i are integers ($\leq k$), and some data are allowed to appear multiple times (e.g., z^5 appears twice in Z^2). As a result, bootstrap sampling can maintain the original distribution of the data and make the generated training sets independent of each other, thereby significantly improving accuracy [51].

Random forest is an ensemble learning algorithm that uses a group of decision trees built by the subtraining sets as weak individual learners of randomly sampled training data. Each decision tree has multiple nodes, and the threshold values of the variables at each node are computationally determined to yield the largest information gain. Generally speaking, the greater the information gain, the greater the "purity improvement" obtained using this feature variable. Since one individual decision tree typically exhibits high variance and tends to overfit, random forest can achieve reduced variance by combining diverse trees, hence yielding a better model overall. Moreover, the out-of-bag data that are not used in the training process for each decision tree can be used to estimate the skill and effectiveness of the trained random forest model. The whole algorithm is conducted in Python using the open-source pandas and scikit-learn machine learning libraries [52].

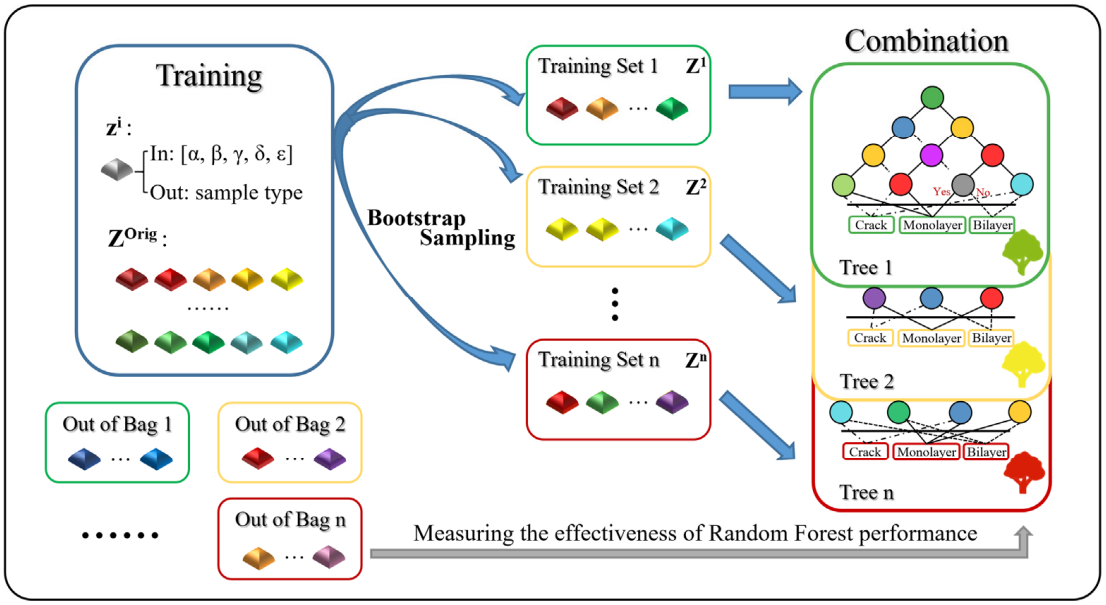

Figure 4. Basic architecture of the learning procedure in the random forest method. Each small square represents a spatial measurement point carrying Raman characteristic information. The subtraining sets from 1 to n are acquired by a bootstrap sampling process, and then decision trees based on these subtraining sets can be built. The out-of-bag data of each tree can be used to estimate the effectiveness of the trained random forest model.

Figure 5a shows the prediction procedure for the random forest method. After generating the whole forest, new samples from different positions carrying the input Raman characteristic information will be judged through the formed decision trees one by one. Since the training sets are different from each other, they may give different judgments for one sample by different trees. The final output result is generated by the democratic majority voting which can acquire a dramatically reduced variance. After the optimal parameter combination is fixed, it can immediately give a classification result for the new input data. The accurate measurement of Raman spectra and the use of a sufficient number of training sets enable the constructed random forest to obtain a relatively high accuracy rate. As shown in Figure 5b–e, we choose different sample areas on several pieces of substrate to test our model, and the classification results via the random forest method are successful in distinguishing the monolayer, crack and bilayer areas. In the central areas of these three regions, random forest can provide reliable results. However, due to the relatively small number of detection points at the boundary regions, the result accuracy is not particularly high. Compared with the Raman mapping of the input variable ε in Figure 5b, we find that the proposed model successfully retains the accuracy of Raman shifts to effectively identify monolayer and bilayer areas. Benefitting from the high-speed computing power of the computer, the random forest algorithm can be easily and continuously strengthened by increasing the amount of data. As the number of learning data of insufficient sample types are increased, the recognition accuracy is expected to be further improved [48].

Furthermore, some dispersed dots in Figure 5b are easily predicted, which would be commonly ignored under optical microscope because of the lower optical contrast. As reported in the previous work [53], several stages were observed during the MoS_2 atomic layer growth. Initially, some small domains were nucleated at random locations on the substrate. Then, the nucleation sites continued to grow and formed boundaries when two or more domains met, resulting in partially continuous films. The as-grown films are predominantly monolayer, with small areas consisting of two or more layers at the preferred nucleation sites [42,53], which explains the manifested bilayer Raman characteristics of these areas located on the monolayer continuous film. This suggests that prediction via the random forest possesses the application potential to view subtle differences through the material's basic features. By using a shorter wavelength laser and a larger numerical aperture objective, it is expected that

the spatial resolution will be further improved, which makes it possible to identify defects that are difficult to find and/or determine only by optical images. A smaller laser spot is ideal for analyzing the microscopic characteristics of the sample. In theory, the higher the spatial resolution, the more precise the micro-area spectral information that can be obtained. The random forest algorithm is not a conservative and fixed tool and it can be flexibly adjusted according to our actual needs to adapt to different scenarios. During the Raman spectra acquisition process, it is better to ensure the stability of the experimental conditions, so as to ensure the accuracy to the greatest extent.

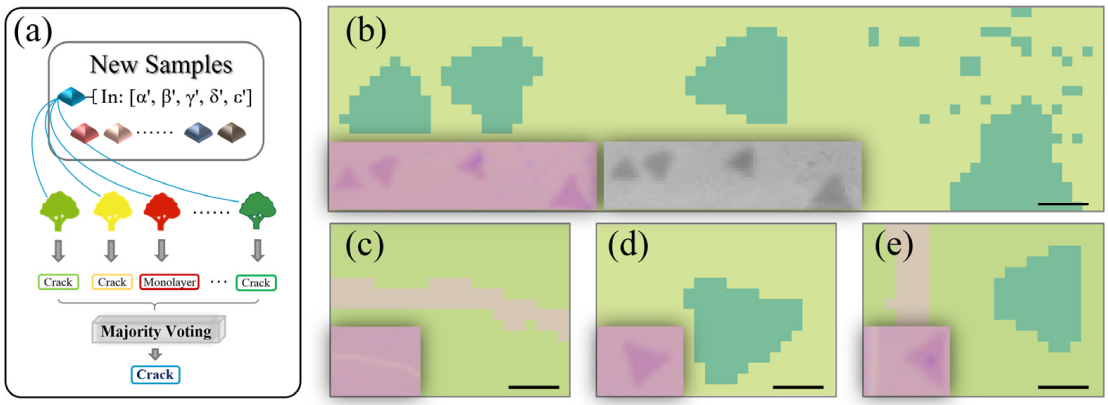

Figure 5. (**a**) Basic architecture of the prediction procedure in the random forest method. The new samples from the untrained data are judged through each tree one by one, and the final output results are acquired by the majority voting process. (**b**–**e**) The predicted pictures for different samples with crack (grown), monolayer (grass green) and bilayer (dark green) areas. The dispersed dots shown in Figure 5b are predicted to be bilayer. The inset figures in Figure 5b show the corresponding optical micrograph (left inset) and Raman mapping of the input variable ε (right inset). The other inset figures show the corresponding optical micrographs. Scale bars indicate 1 μm.

To further demonstrate the performance of our built random forest model, the receiver operating characteristic (ROC) and precision-recall (PR) curves are analyzed for crack/others and bilayer/others, respectively [25]. We use pre-labeled new data that does not appear as the testing set to calculate the accuracy, which is used to evaluate the performance of the trained model. The ROC curve is a popular method for accuracy assessment because it is comprehensive, understandable and visually attractive [54,55]. This method uses the area under the curve (AUC) for quantitative assessment, which plots 1-specificity on the x-axis against sensitivity on the y-axis. The range of the AUC varies from 0.5 to 1.0 and a perfect predictor gives an AUC score of 1, while a predictor that makes random guesses like coin tossing has an AUC score of 0.5. While the PR curve shows the trade-off between precision and recall for different thresholds. A system with high recall but low precision will return many results, but most of its predicted labels are incorrect in comparison to the training labels. However, a system with high precision but low recall is just the opposite, returning very few results, but most of the predicted labels are correct in comparison to the training labels. An ideal system with high precision and high recall will return many results, with all results labeled correctly.

As shown in Figure 6a, when using only the Raman frequency difference between the E^1_{2g} and A_{1g} peaks, the AUC is 0.9891 for bilayer/others but is only 0.7543 for crack/others, which means that the Raman frequency difference can recognize bilayer samples extremely well but is not good for the cracks. Uniformly, the PR curves in Figure 6b also prove this point, in which the average precision (AP) values are 0.9895 and 0.7167 for bilayer/others and crack/others, respectively. However, when taking all the input variables (α, β, γ, δ and ε) into consideration, the numerical values of AUC and AP are 0.9852 and 0.9867 for crack, and 0.9902 and 0.9914 for bilayer, which means that this technique based on the random forest method can successfully characterize and confirm the monolayer, crack and bilayer areas at the same time.

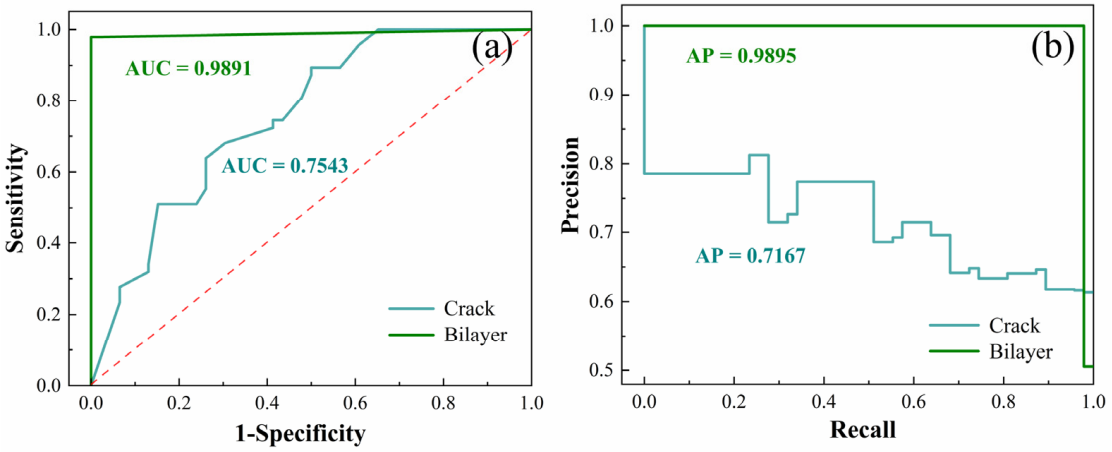

Figure 6. (**a**) Receiver operating characteristic (ROC) curves for the crack and bilayer identification only use the characteristic of the Raman frequency difference. Cyan and green curves show the ROC for two-class classifications of crack/others and bilayer/others, respectively. The red line corresponds to the situation of a random guess. (**b**) Precision-recall (PR) curves for the crack and bilayer identification only use the characteristic of the Raman frequency difference.

Generally, unsupervised techniques, such as CNN and U-Net, are more robust, but they require immense training data and higher computing resources. However, supervised machine learning, like SVM and random forest, deals with prelabeled training data, which not only reduces the computational expense and time but also achieves high accuracy especially when the dimension of the feature vector is not very large [56]. Compared with using the optical contrast via pixel intensities of red, green and blue, the Raman features can directly describe and reflect the intrinsic characteristic differences in materials, which means that even a reduced number of training sets can also help us to get relatively accurate results in a short time. Among the machine learning algorithms, the random forest algorithm has been proven to have unique potential in processing spectral data, with benefits due to its high accuracy and strong resistance to over-fitting. The introduction of machine learning makes it possible to continuously learn in spectroscopy research.

This work is a preliminary step and an attempt to combine machine learning methods with traditional Raman spectroscopy, but it has great migration possibilities for other 2D materials and even bulk materials. Since the Raman spectra of various defects in imperfect 2D materials exhibit different changes, we expect that this method will play a greater role in the characterization of more complex material property control engineering, such as doping, oxidation, mechanical deformations, etc. Machine learning algorithms can be used to build databases under different equipment and experimental conditions, which can better help us analyze and compare experimental data. At the same time, the abundant information in the spectra makes it possible for machine learning to solve different problems by extracting different multi-dimensional variables. The power of machine learning is rapidly transforming modern science, and we can anticipate more exciting results stemming from this interplay between machine learning and the physical sciences [57]. Nowadays, the confluence of many traditional and emerging disciplines, for example, nano-manufacturing, big data technology, computer science and artificial intelligence, is expected to lead the trend in the theoretical and experimental advances in exploring 2D materials, and usher in abundant research opportunities for developing novel 2D devices and systems.

4. Conclusions

In this study, we demonstrated an effective method based on a random forest algorithm to classify monolayer MoS_2 continuous film, random crack and bilayer areas from the variables extracted from Raman spectra. The random forest method was used to analyze multiple Raman features to identify samples, solving the problem of ineffective identification of samples by just one specific variable. It can

successfully determine the importance of a certain characteristic variable and some dispersed defect dots can also be predicted, which would commonly be ignored under an optical microscope. By taking peak intensity and frequency information into consideration at the same time, a high accuracy rate is obtained. The method developed in this work can also be used for other 2D materials and can provide a valuable reference for material characterization in several fields.

Author Contributions: Conceptualization, Y.M., N.D., L.W. and J.W.; methodology, Y.M. and L.W.; software, Y.M.; validation, Y.M., X.C., H.W. and Z.W.; data curation, Y.M.; writing—original draft preparation, Y.M.; writing—review and editing, Y.M., N.D. and J.W.; supervision, N.D. and J.W.; project administration, J.W.; funding acquisition, I.M.K. and J.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China (NSFC) (nos. 11904375, 61975221) and the Strategic Priority Research Program of the Chinese Academy of Sciences (CAS) (nos. XDB16030700, XDB43010303).

Acknowledgments: The authors acknowledge financial support from Shanghai Science and Technology International Cooperation Fund (No. 19520743900) and the CAS President's International Fellowship Initiative (No. 2021VTB0003).

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

References

1. Tan, C.; Cao, X.; Wu, X.-J.; He, Q.; Yang, J.; Zhang, X.; Chen, J.; Zhao, W.; Han, S.; Nam, G.-H.; et al. Recent Advances in Ultrathin Two-Dimensional Nanomaterials. *Chem. Rev.* **2017**, *117*, 6225–6331. [CrossRef] [PubMed]
2. Wang, G.; Chernikov, A.; Glazov, M.M.; Heinz, T.F.; Marie, X.; Amand, T.; Urbaszek, B. Colloquium: Excitons in atomically thin transition metal dichalcogenides. *Rev. Mod. Phys.* **2018**, *90*, 021001. [CrossRef]
3. Mueller, T.; Malic, E. Exciton physics and device application of two-dimensional transition metal dichalcogenide semiconductors. *npj 2D Mater. Appl.* **2018**, *2*, 29. [CrossRef]
4. Zhan, Y.; Liu, Z.; Najmaei, S.; Ajayan, P.M.; Lou, J. Large-Area Vapor-Phase Growth and Characterization of MoS_2 Atomic Layers on a SiO_2 Substrate. *Small* **2012**, *8*, 966–971. [CrossRef]
5. Zhang, Y.; Zhang, L.; Zhou, C. Review of Chemical Vapor Deposition of Graphene and Related Applications. *Acc. Chem. Res.* **2013**, *46*, 2329–2339. [CrossRef] [PubMed]
6. Wu, Y.; Li, Z.; Liao, W.; Jia, Y.; Shi, Z.; Huang, Z.; Yu, W.; Sun, X.; Liu, X.; Li, D. Monolithic integration of MoS_2-based visible detectors and GaN-based UV detectors. *Photonics Res.* **2019**, *7*, 1127–1133. [CrossRef]
7. Lu, R.D.; Wang, Y.G.; Wang, J.; Ren, W.; Li, L.; Liu, S.C.; Chen, Z.D.; Li, Y.F.; Wang, H.Y.; Fu, F.X. Soliton and bound-state soliton mode-locked fiber laser based on a MoS_2/fluorine mica Langmuir-Blodgett film saturable absorber. *Photonics Res.* **2019**, *7*, 431–436. [CrossRef]
8. Liu, W.; Liu, M.; Han, H.; Fang, S.; Teng, H.; Lei, M.; Wei, Z. Nonlinear optical properties of WSe_2 and $MoSe_2$ films and their applications in passively Q-switched erbium doped fiber lasers. *Photonics Res.* **2018**, *6*, C15–C21. [CrossRef]
9. Rhodes, D.; Chae, S.H.; Ribeiro-Palau, R.; Hone, J. Disorder in van der Waals heterostructures of 2D materials. *Nat. Mater.* **2019**, *18*, 541–549. [CrossRef]
10. Cai, Z.; Liu, B.; Zou, X.; Cheng, H.-M. Chemical Vapor Deposition Growth and Applications of Two-Dimensional Materials and Their Heterostructures. *Chem. Rev.* **2018**, *118*, 6091–6133. [CrossRef] [PubMed]
11. Splendiani, A.; Sun, L.; Zhang, Y.; Li, T.; Kim, J.; Chim, C.-Y.; Galli, G.; Wang, F. Emerging Photoluminescence in Monolayer MoS_2. *Nano Lett.* **2010**, *10*, 1271–1275. [CrossRef] [PubMed]
12. Eda, G.; Yamaguchi, H.; Voiry, D.; Fujita, T.; Chen, M.; Chhowalla, M. Photoluminescence from Chemically Exfoliated MoS_2. *Nano Lett.* **2011**, *11*, 5111–5116. [CrossRef]
13. Li, H.; Wu, J.; Huang, X.; Lu, G.; Yang, J.; Lu, X.; Xiong, Q.; Zhang, H. Rapid and Reliable Thickness Identification of Two-Dimensional Nanosheets Using Optical Microscopy. *ACS Nano* **2013**, *7*, 10344–10353. [CrossRef]
14. Wang, H.-C.; Huang, S.-W.; Yang, J.-M.; Wu, G.-H.; Hsieh, Y.-P.; Feng, S.-W.; Lee, M.K.; Kuo, C.-T. Large-area few-layered graphene film determination by multispectral imaging microscopy. *Nanoscale* **2015**, *7*, 9033–9039. [CrossRef]

15. Li, Y.; Dong, N.; Zhang, S.; Wang, K.; Zhang, L.; Wang, J. Optical identification of layered MoS_2 via the characteristic matrix method. *Nanoscale* **2016**, *8*, 1210–1215. [CrossRef]
16. Niu, Y.; Gonzalez-Abad, S.; Frisenda, R.; Marauhn, P.; Drüppel, M.; Gant, P.; Schmidt, R.; Taghavi, N.S.; Barcons, D.; Molina-Mendoza, A.J.; et al. Thickness-Dependent Differential Reflectance Spectra of Monolayer and Few-Layer MoS_2, $MoSe_2$, WS_2 and WSe_2. *Nanomater.* **2018**, *8*, 725. [CrossRef] [PubMed]
17. Wang, Y.; Zhang, L.; Su, C.; Xiao, H.; Lv, S.; Zhang, F.; Sui, Q.; Jia, L.; Jiang, M. Direct Observation of Monolayer MoS_2 Prepared by CVD Using In-Situ Differential Reflectance Spectroscopy. *Nanomater.* **2019**, *9*, 1640. [CrossRef] [PubMed]
18. Lee, C.; Yan, H.; Brus, L.E.; Heinz, T.F.; Hone, J.; Ryu, S. Anomalous Lattice Vibrations of Single- and Few-Layer MoS_2. *ACS Nano* **2010**, *4*, 2695–2700. [CrossRef] [PubMed]
19. Li, H.; Zhang, Q.; Yap, C.C.R.; Tay, B.K.; Edwin, T.H.T.; Olivier, A.; Baillargeat, D. From Bulk to Monolayer MoS_2: Evolution of Raman Scattering. *Adv. Funct. Mater.* **2012**, *22*, 1385–1390. [CrossRef]
20. Yuan, S.; Liu, L.; Wang, Z.; Xi, N. *AFM-Based Observation and Robotic Nano-Manipulation. AFM-Based Observation and Robotic Nano-Manipulation*; Springer: Singapore, 2020.
21. Li, X.-L.; Han, W.-P.; Wu, J.-B.; Qiao, X.-F.; Zhang, J.; Tan, P.-H. Layer-Number Dependent Optical Properties of 2D Materials and Their Application for Thickness Determination. *Adv. Funct. Mater.* **2017**, *27*, 1604468. [CrossRef]
22. Hornett, S.M.; Stantchev, R.I.; Vardaki, M.; Beckerleg, C.; Hendry, E. Subwavelength Terahertz Imaging of Graphene Photoconductivity. *Nano Lett.* **2016**, *16*, 7019–7024. [CrossRef]
23. Lin, X.; Si, Z.; Fu, W.; Yang, J.; Guo, S.; Cao, Y.; Zhang, J.; Wang, X.; Liu, P.; Jiang, K.; et al. Intelligent identification of two-dimensional nanostructures by machine-learning optical microscopy. *Nano Res.* **2018**, *11*, 6316–6324. [CrossRef]
24. Masubuchi, S.; Machida, T. Classifying optical microscope images of exfoliated graphene flakes by data-driven machine learning. *npj 2D Mater. Appl.* **2019**, *3*, 4. [CrossRef]
25. Saito, Y.; Shin, K.; Terayama, K.; Desai, S.; Onga, M.; Nakagawa, Y.; Itahashi, Y.M.; Iwasa, Y.; Yamada, M.; Tsuda, K. Deep-learning-based quality filtering of mechanically exfoliated 2D crystals. *npj Comput. Mater.* **2019**, *5*, 1–6. [CrossRef]
26. Masubuchi, S.; Watanabe, E.; Seo, Y.; Okazaki, S.; Sasagawa, T.; Watanabe, K.; Taniguchi, T.; Machida, T. Deep-learning-based image segmentation integrated with optical microscopy for automatically searching for two-dimensional materials. *npj 2D Mater. Appl.* **2020**, *4*, 1–9. [CrossRef]
27. Han, B.; Lin, Y.; Yang, Y.; Mao, N.; Li, W.; Wang, H.; Yasuda, K.; Wang, X.; Fatemi, V.; Zhou, L.; et al. Deep-Learning-Enabled Fast Optical Identification and Characterization of 2D Materials. *Adv. Mater.* **2020**, *32*, e2000953. [CrossRef]
28. Millard, T.S.; Genco, A.; Alexeev, E.M.; Randerson, S.; Ahn, S.; Jang, A.-R.; Shin, H.S.; Tartakovskii, A.I. Large area chemical vapour deposition grown transition metal dichalcogenide monolayers automatically characterized through photoluminescence imaging. *npj 2D Mater. Appl.* **2020**, *4*, 12. [CrossRef]
29. Nolen, C.M.; Denina, G.; Teweldebrhan, D.; Bhanu, B.; Balandin, A.A. High-Throughput Large-Area Automated Identification and Quality Control of Graphene and Few-Layer Graphene Films. *ACS Nano* **2011**, *5*, 914–922. [CrossRef]
30. Dhakal, K.P.; Duong, D.L.; Lee, J.; Nam, H.; Kim, M.; Kan, M.; Lee, Y.H.; Kim, J. Confocal absorption spectral imaging of MoS_2: Optical transitions depending on the atomic thickness of intrinsic and chemically doped MoS_2. *Nanoscale* **2014**, *6*, 13028–13035. [CrossRef]
31. Desai, S.B.; Madhvapathy, S.R.; Amani, M.; Kiriya, D.; Hettick, M.; Tosun, M.; Zhou, Y.; Dubey, M.; Ager, J.W.; Chrzan, D.; et al. Gold-Mediated Exfoliation of Ultralarge Optoelectronically-Perfect Monolayers. *Adv. Mater.* **2016**, *28*, 4053–4058. [CrossRef] [PubMed]
32. Zhang, X.; Qiao, X.-F.; Shi, W.; Wu, J.-B.; Jiang, D.-S.; Tan, P.-H. Phonon and Raman scattering of two-dimensional transition metal dichalcogenides from monolayer, multilayer to bulk material. *Chem. Soc. Rev.* **2015**, *44*, 2757–2785. [CrossRef] [PubMed]
33. Dai, X.; Zhang, X.; Kislyakov, I.M.; Wang, L.; Huang, J.; Zhang, S.; Dong, N.; Wang, J. Enhanced two-photon absorption and two-photon luminescence in monolayer MoS_2 and WS_2 by defect repairing. *Opt. Express* **2019**, *27*, 13744–13753. [CrossRef]

34. Li, Y.; Dong, N.; Zhang, S.; Zhang, X.; Feng, Y.; Wang, K.; Zhang, L.; Wang, J. Giant two-photon absorption in monolayer MoS_2. *Laser Photonics Rev.* **2015**, *9*, 427–434. [CrossRef]
35. Urban, F.; Passacantando, M.; Giubileo, F.; Iemmo, L.; Di Bartolomeo, A. Transport and Field Emission Properties of MoS_2 Bilayers. *Nanomater.* **2018**, *8*, 151. [CrossRef]
36. Xie, Y.; Zhang, S.; Li, Y.; Dong, N.; Zhang, X.; Wang, L.; Liu, W.; Kislyakov, I.M.; Nunzi, J.-M.; Qi, H.; et al. Layer-modulated two-photon absorption in MoS_2: Probing the shift of the excitonic dark state and band-edge. *Photonics Res.* **2019**, *7*, 762–770. [CrossRef]
37. Vaknin, Y.; Dagan, R.; Rosenwaks, Y. Pinch-Off Formation in Monolayer and Multilayers MoS_2 Field-Effect Transistors. *Nanomaterials* **2019**, *9*, 882. [CrossRef]
38. Hao, S.; Yang, B.; Gao, Y. Orientation-specific transgranular fracture behavior of CVD-grown monolayer MoS_2 single crystal. *Appl. Phys. Lett.* **2017**, *110*, 153105. [CrossRef]
39. Wang, K.; Wang, J.; Fan, J.; Lotya, M.; O'Neill, A.; Fox, D.; Feng, Y.; Zhang, X.; Jiang, B.; Zhao, Q.; et al. Ultrafast Saturable Absorption of Two-Dimensional MoS_2 Nanosheets. *ACS Nano* **2013**, *7*, 9260–9267. [CrossRef]
40. Yoon, D.; Moon, H.; Son, Y.-W.; Choi, J.S.; Park, B.H.; Cha, Y.H.; Kim, Y.D.; Cheong, H. Interference effect on Raman spectrum of graphene on SiO_2/Si. *Phys. Rev. B* **2009**, *80*, 125422. [CrossRef]
41. Li, X.-L.; Qiao, X.-F.; Han, W.-P.; Zhang, X.; Tan, Q.-H.; Chen, T.; Tan, P.-H. Determining layer number of two-dimensional flakes of transition-metal dichalcogenides by the Raman intensity from substrates. *Nanotechnology* **2016**, *27*, 145704. [CrossRef]
42. Lee, Y.-H.; Zhang, X.; Zhang, W.; Chang, M.-T.; Lin, C.-T.; Chang, K.-D.; Yu, Y.-C.; Wang, J.T.-W.; Chang, C.-S.; Li, L.-J.; et al. Synthesis of Large-Area MoS_2 Atomic Layers with Chemical Vapor Deposition. *Adv. Mater.* **2012**, *24*, 2320–2325. [CrossRef]
43. Jain, A.K. Data clustering: 50 years beyond K-means. *Pattern Recognit. Lett.* **2010**, *31*, 651–666. [CrossRef]
44. Brown, D.B.; Shen, W.; Li, X.; Xiao, K.; Geohegan, D.B.; Kumar, S. Spatial Mapping of Thermal Boundary Conductance at Metal–Molybdenum Diselenide Interfaces. *ACS Appl. Mater. Interfaces* **2019**, *11*, 14418–14426. [CrossRef]
45. Ferrari, A.C. Raman spectroscopy of graphene and graphite: Disorder, electron–phonon coupling, doping and nonadiabatic effects. *Solid State Commun.* **2007**, *143*, 47–57. [CrossRef]
46. Ferrari, A.C.; Meyer, J.C.; Scardaci, V.; Casiraghi, C.; Lazzeri, M.; Mauri, F.; Piscanec, S.; Jiang, D.; Novoselov, K.S.; Roth, S.; et al. Raman Spectrum of Graphene and Graphene Layers. *Phys. Rev. Lett.* **2006**, *97*, 187401. [CrossRef]
47. Zhang, X.; Han, W.P.; Wu, J.B.; Milana, S.; Lu, Y.; Li, Q.Q.; Ferrari, A.C.; Tan, P.H. Raman spectroscopy of shear and layer breathing modes in multilayer MoS_2. *Phys. Rev. B* **2013**, *87*, 115413. [CrossRef]
48. Tanaka, K.; Hachiya, K.; Zhang, W.; Matsuda, K.; Miyauchi, Y. Machine-Learning Analysis to Predict the Exciton Valley Polarization Landscape of 2D Semiconductors. *ACS Nano* **2019**, *13*, 12687–12693. [CrossRef]
49. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
50. Cutler, D.R.; Edwards, T.C., Jr.; Beard, K.H.; Cutler, A.; Hess, K.T.; Gibson, J.; Lawler, J.J. Random forests for classification in ecology. *Ecology* **2007**, *88*, 2783–2792. [CrossRef]
51. Breiman, L. Bagging predictors. *Mach. Learn.* **1996**, *24*, 123–140. [CrossRef]
52. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830. [CrossRef]
53. Najmaei, S.; Liu, Z.; Zhou, W.; Zou, X.; Shi, G.; Lei, S.; Yakobson, B.I.; Idrobo, J.-C.; Ajayan, P.M.; Lou, J. Vapour phase growth and grain boundary structure of molybdenum disulphide atomic layers. *Nat. Mater.* **2013**, *12*, 754–759. [CrossRef]
54. Chen, W.; Li, Y.; Xue, W.; Shahabi, H.; Li, S.; Hong, H.; Wang, X.; Bian, H.; Zhang, S.; Pradhan, B.; et al. Modeling flood susceptibility using data-driven approaches of naïve Bayes tree, alternating decision tree, and random forest methods. *Sci. Total Environ.* **2020**, *701*, 134979. [CrossRef]
55. Shi, J.; Wang, Y.; Chen, T.; Xu, D.; Zhao, H.; Chen, L.; Yan, C.; Tang, L.; He, Y.; Feng, H.; et al. Automatic evaluation of traumatic brain injury based on terahertz imaging with machine learning. *Opt. Express* **2018**, *26*, 6371–6381. [CrossRef]

56. Yang, J.; Yao, H. Automated identification and characterization of two-dimensional materials via machine learning-based processing of optical microscope images. *Extrem. Mech. Lett.* **2020**, *39*, 100771. [CrossRef]
57. Carleo, G.; Cirac, I.; Cranmer, K.; Daudet, L.; Schuld, M.; Tishby, N.; Vogt-Maranto, L.; Zdeborová, L. Machine learning and the physical sciences. *Rev. Mod. Phys.* **2019**, *91*, 045002. [CrossRef]