Abstract—Medicinal plants are a primary source of disease treatment in many countries. As most are edible however, consumption of the wrong herbal plants can have serious consequences and even lead to death. Automatic accurate recognition of plant species to help users who do not have specialist knowledge of herbal plants is thus a desirable aim. Several automatic medicinal plant identification systems have been proposed, though most are significantly constrained either in the small number of species or in requiring manual image segmentation of plant leaves. This means they are captured on a plain background rather than being readily identified in their natural surroundings, which often involve complex and noisy backgrounds. While deep learning (DL) based methods have made considerable strides in recent times, their potential has not always been maximised because they are trained with samples which are not always fully representative of the intra-class and inter-class differences between the plant species concerned. This paper addresses this challenge by incorporating mutual information into a Convolutional Neural Network (CNN) model to select samples for the training, validation, and testing sets based on a similarity measure. A critical comparative evaluation of this new CNN medicinal plant classification model incorporating a mutual information guided training (MIGT) algorithm for sample selection, corroborates the superior classification performance achieved for the VNPlant-200 dataset, with an average accuracy of more than 97%, while the precision and recall values are also consistently above 97%. This is significantly better than existing CNN classification methods for this dataset as it crucially means false positive rates are substantially lower thus affording improved identification reliability.

Index Terms—Medicinal plants, Natural Environment, Convolutional Neural Networks, Mutual Information, training dataset.

I. INTRODUCTION

Plants play an essential role for human life to survive on the earth, in terms of food, fibre, shelter, fuel and medicine [1]. More than 60% of the population in developing countries rely on traditional therapeutic herbal plants as a remedy for various human diseases [2]. In this context, it is thus desirable to have an automated medicinal plant recognition system which can help users, who neither have an in-depth knowledge nor specialist botany training to be able to accurately identify medicinal plants in their natural surroundings from a variety of images. Accurate identification of such medicinal plants is a key user requirement as consuming wrongly identified plants can cause harm or in the worst case, even the death [3]. Taxonomists are seeking ever more effective and reliable ways of plant species identification, including adopting computer vision, pattern recognition and artificial intelligence techniques to improve the accuracy and efficiency of medicinal plant classification. Image-based methods are an especially fertile approach for plant identification, given the ubiquitous nature of mobile phones which readily can capture images and then analyse them via a suitable plant recognition application to either identify a species or obtain a likely list of species in instances where a single match is not feasible [2], [4].

Automatic medicinal plant species identification, however, faces many logistical and technical challenges that require a robust solution. For example, plants are often blighted by disease, which commonly affects the surface of their leaves leading to pigmentation, discoloring or even change of the leaf shape, which means this key morphological feature is difficult to distinguish. Leaves also have an inherent 3D shape that is not be fully represented in 2D images [5].

Only a small collection of image datasets is available which consider plants in their natural surroundings such as the BJFU100 dataset [6], from which some sample images are shown in Fig. 1(a). Traditionally plant images have been taken by high-quality cameras in laboratory settings rather than in their natural habitat, and usually on a plain background isolated from petioles. In contrast, image captured in the wild invariably suffer from illumination and shading effects, occlusions, background clutter, disease, deformity, and other assorted visual artefacts. The resulting image quality is generally not so high due to camera and/or hand motion. For these reasons, research has mainly focused on scanned and segmented plant images against a white background [7] as illustrated in Fig. 1(b) [8]. This does not however, reflect the real world as images are captured under controlled lighting conditions and many of the aforesaid extrinsic factors are excluded, so the accuracy of a plant classifier will inevitably be higher. While the human visual system instinctively deals
with many of these external factors, a computational model for automated categorisation of plant from natural images remains a significant challenge.

While some classification solutions for medicinal plants with complex backgrounds have been proposed [9], [10] the number of species evaluated is generally small, so the critical challenge of accurately classifying plants with large variations in intra-class and small variations in inter-class [5] is not considered. In such small datasets, classification accuracy tends to be higher because subtle feature differences or similarities between different plant species are not faithfully reflected.

Although some plants can be identified based on the shape of leaves, others exhibit considerable degrees of similarity despite being from completely different species and can only be identified by non-morphological features which are often hidden.

Botanist and taxonomists have observed that plants belonging to the same species display distinct variations in their morphological characteristics for various reasons including genetic mutations, environmental and climatic factors like nutrition, levels of sunlight, geographical heterogeneity in the soil type, and seasonal variations in temperature, humidity, and rainfall [5], [11]. Fig. 2 provides an example of a plant from the same species with demonstrably different leaf and flower characteristics, as well as colour and shape.

Current literature mainly evaluates the accuracy performance of plant identification systems [12], [13] though by only using this metric in medicinal applications, it importantly fails to penalise false positives, which can have serious repercussions. Precision, as the ratio of correctly detected positive cases to all detected cases to be positive, is a critical metric for medicinal plants, as results will reveal that even with an accuracy of 89% the corresponding precision may only be around 30% for certain species. Thus, in critically evaluating the results of our method, as well as accuracy, precision, recall (estimation of affectability) and F1-scores are considered [14].

In larger medicinal plant datasets of images taken in their natural environment, under certain conditions, intra-class species variations can be significant, while inter-class variability can be small. The corollary is that a very high misclassification rate can ensue. To address this problem, this paper presents a novel samples selection approach to improve the testing, training, and validation steps of a CNN model. Critical evaluation results corroborate that by incorporating mutual information (MI) as a similarity measure, the model takes broader cognisance of intra-class variations in key features of specific species in the training step, leading to a more accurate and dependable classification performance. This provides both the layperson and specialists with greater confidence in reliably identifying species in their natural environment.

The remainder of the paper is organised as follows: Section II reviews existing image processing and DL techniques used for the classification of medicinal plants. Section III details the MIGT algorithm, while Section IV providing a critical comparative performance evaluation of using MIGT in a CNN model. Finally, Section V makes some concluding and future work comments.

II. RELATED WORKS

Research into plant species classification has focused on a variety of technologies, with DL methods increasingly offering promising results and high accuracy over classical classification techniques [6], [15]. Reference [3] proposed medicinal plant classification using a DL network inspired from AlexNet for feature extraction, with SoftMax and Support Vector Machine (SVM) applied for classification to achieve an accuracy of 96.76%, though only a small number (40) of species were analysed and only the plant leaves scanned, with the image backgrounds being segmented manually, so each leaf was against a plain background. Accuracy was used for evaluating the classifier performance, though other important metrics such as precision, recall, and F1-score were not considered. An alternative approach proposed by [12], synthesised a dataset containing ten species of Indian medicinal plants with 20950 images and employing augmentation techniques [16] and different CNN models. The Mobilenet model with 500 steps was the best classified giving a 98.2% accuracy. While their focus was on plants in the natural environment, the number of species was limited and classification results evaluated using only accuracy, so false positive rates i.e., the precision, recall and F1-scores were not considered. Paulson and Ravishankar [13] presented a dataset of 64 species of medicinal plants with 1000 samples for each species. They had a comparative study.
between pre-trained CNN models, namely Visual Geometry Group Network (VGG)16 and VGG19, and their proposed CNN. The experimental results show that VGG16 achieved the highest classification accuracy of 97.8%. VGG19 achieved 97.6% accuracy, and basic CNN achieved 95.8% accuracy, though again they only consider images with plain background and use only accuracy rather than the more holistic confusion matrix metrics in the performance evaluation.

Quoc and Hoang [17] introduced a benchmark Vietnamese VNPlant-200 dataset with 200 species of medicinal plants, and which represents the first large, public, and multi-class collection of such plants. This dataset is more challenging and noisier than existing datasets as many leaves appear together in a single image, and it also contains the background such as soil, tree bark and flower. The Speed-Up Robust Features (SURF) and Scale Invariant Feature Transform (SIFT) algorithms are employed to extract features associated with the Bag-of-Words (BOW) technique to create feature vectors, and Random Forest (RF) for classification. The SIFT descriptor detects points of interest to describe many different objects in an image. The SURF descriptor is inspired by the SIFT descriptor but is a faster algorithm and is more robust against different image transformations. The feature extraction from SURF or SIFT is combined with BOW using K-means with \( K = 100 \), and the RF classifier has the number of trees \( n = 1000 \). The dataset has two different pixel resolutions, namely 256\( \times 256 \) and 512\( \times 512 \), obtaining the best accuracy of 37.4% with SIFT method for larger sized images. They subsequently presented comparative findings for different CNN architectures on their dataset, using six CNN models (VGG16, ResNet50, Inceptionv3, DenseNet121, Xception, and MobileNetV2). The results show that the Xception model reached 88.26% and outperforms other models [15], though a significant number of images were misclassified. Furthermore, the authors only measured success based on average accuracy, so the false positive and negatives rates were not considered.

Medicinal plant identification has principally focused on datasets comprising leaf images on simple backgrounds with only a very limited number of datasets addressing plants in their natural environment. Furthermore, while accuracy is a widely adopted metric for performance evaluation, it alone is not sufficiently robust to reliably classify medicinal plant images because low false positive rates are a significant metric to correctly identifying species. In this context, other confusion matrix metrics like precision and recall must be included to provide more holistic insights into the classification performance of CNN-based models for medicinal plants.

### III. THE MUTUAL INFORMATION GUIDED TRAINING (MIGT) ALGORITHM

#### A. Dataset

The VNPlant-200 dataset was chosen for this investigation as it is the most comprehensive and challenging collection of medicinal plants available, with all images being captured in their natural environment. Fig. 3 shows some samples from this dataset, which contains 20,000 images of 200 different medicinal plant species, with each species containing 100 separate images [17]. In such unconstrained conditions, there are inevitably considerable variations in illumination, rotation, scale for certain species, so every image has been labelled manually by botanists and domain experts. Despite this, many species can still be difficult to accurately identify because leaves sometimes appear to overlap with each other, and there frequent occlusions and complex backgrounds involving soil, tree bark, and flowers.

#### B. Mutual Information (MI)

In analysing the experimental results in [15], it is apparent the classification performance for certain species is notably lower than the overall accuracy. These classification errors are caused by the CNN model being trained by image samples that fail to reflect broader intra-class variations for the species. The samples used for training, testing and validation are usually all randomly chosen from the dataset, with Figs. 4(a), 4(b) and 4(c) respectively illustrating examples of these. It is evident that colour features like dull green leaves dominate the testing set for this species, though this characteristic is not noticeable in either the training or validation sets, so predictably the classifier performance drops for these images. The training samples thus need to be more representative of each species class, with similar valid arguments for the testing and validation datasets to avoid the introduction of classification bias. This provided the motivation to investigate embedding a suitable similarity measure method into the CNN model for selecting appropriate intra-class species training, testing and validation samples that better reflect the extent of intra-class differences for each species.

In selecting training samples, images with various distinguishing characteristics of a particular species class need to be included in the training set. A similarity measure (SM) method, which quantifies the correspondence between two images within a class, is thus applied to determine how closely the image samples are related to guide the selection of samples which best reflect the distinguishing appearances of that class in the training set. Various SM methods, including sum-of-square distance (SSD), mean square distance (MSD), (normalized) cross correlation (CC) and MI have been considered.
The pixel intensity based methods such as SSD and MSD are unsuitable for this application as the complex background, colour and size of plants are not directly comparable. CC is useful for searching a long signal for a shorter known feature, but it is not very effective for comparing two full images [18].

MI is an entropy based method that measures how much information one variable is conveyed in another. For images with complex backgrounds and the object of interest being various colours, shapes and sizes, MI is an especially attractive option. Furthermore, MI values converge to tight error bounds to reflect the true relatedness [19].

More specifically, MI is a similarity measure that quantifies the relationship between two random variables that are sampled simultaneously. It measures how much average information is communicated, in one random variable about another. It, therefore, looks at the amount of uncertainty that is lost from one variable when the other is known [20]. MI can be expressed as the difference between two entropy terms.

\[ I(X; Y) = H(X) - H(X | Y) = H(Y) - H(Y | X) \]  

If \( X \) and \( Y \) are independent of each other, their MI is zero, i.e. \( I(X;Y)= 0 \) [21].

In the MITG algorithm, MI guides the selection of samples in deciding the training, testing and validation sets. The relevant pseudo-code and block diagram of the MITG method are presented in Algorithm. 1 and Fig. 5 respectively.

As each bin contains similar MI values, they represent a distinguish appearance type of each species class. The rationale for setting the threshold at 10 samples is that with 50\% being selected for training (and 40\% for testing and 10\% for validation), 5 images is the minimum number which can pragmatically be used to train the CNN model to recognise a particular plant characteristic. The VNPlant-200 dataset images are the input, and the MIGT algorithm returns the training, testing and validation sets as the output. The main loop (Lines 6 to 27) considers each individual species class.
and calculates the corresponding MI values for each image in that species (Line 7). Histogram analysis is then conducted (Line 8) with the bin size being reduced (Lines 8 to 29) until either the lengths of all the bins are greater than the minimum length, which is either set to 10 in this experiment, or when only a single bin remains.

If it is the former, half the content of each bin is allocated to the training set, 40% to the testing and 10% to the validation set (Lines 10 to 17). In the case where only 1 bin remains, all MI values are ranked in ascending order before then being equi-divided into 3 portion (Lines 19 to 23), with each portion containing similar MI values reflecting the distinguishing appearance type.

C. The Convolutional Neural Network (CNN) model

Various CNN models have been evaluated upon the VNPlant-200 dataset images [15] including VGG16, Resnet50, InceptionV3, DenseNet121, Xception (Extreme Inception) and MobileNet, with the Xception model achieving the highest accuracy of over 88%. For this reason, this model was chosen for this investigation.

It has 36 convolutional layers forming the feature extraction base of the network, which are structured into 14 modules, all of which have linear residual connections, except for the first and last modules [22]. As discussed in Section III B, MI measures the similarity between the sample images of each species. These similarity values are then used to guide the selection of train, validation, and test sets such that each set contain distinguished representative samples of each species. The images are then augmented and resized as described by [15] to fit the Xception model’s inputs. Finally, the Xception model classifies each species image.

IV. RESULTS DISCUSSION

Google Colab is used as an experimental testbed platform as it enables DL model development in a web browser with no configuration required. The Xception model is applied to both training and testing, while the large VNPlant-200 dataset has been used for comparative benchmarking purposes.

To ensure an equitable comparison in critically evaluating the classification performance of the MIGT algorithm in the CNN model, the same test set as [15] was employed. Tables I to V summarise the corresponding classification performance of the MIGT algorithm, with Table I revealing a significant overall reduction in the number of misclassifications of nearly one third.

### TABLE I

<table>
<thead>
<tr>
<th>Method</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>Randomly Selection [15]</td>
<td>787</td>
</tr>
<tr>
<td>MIGT</td>
<td>220</td>
</tr>
</tbody>
</table>

While the MIGT algorithm has considerably reduced the misclassification rate, some species are still erroneously classified, including, Croton oblongifolius, Artocarpus heterophyllus, and Plukenetia volubilis. These misclassifications are caused by either a high similarity with another species from the dataset or poor image quality. As mentioned previously, certain species have only small inter-class differences so when there are insufficient samples to discriminate between them, the classifier cannot be suitably trained to reliably distinguish between such species.

Fig. 6 demonstrates four examples of species with relatively high misclassifications. Figs. 6 (a) and 6 (b), have strikingly similar leaves yet are different species so can easily be misclassified, while Figs. 6 (c) and 6 (d) show examples of highly similar species samples from the dataset. Interestingly, the leaves in the latter examples are very small which compounds the challenge to accurately discern between the two species.

Table. II benchmarks the classification performance of the MIGT method using the confusion matrix metrics of accuracy, precision, recall and F1-score [23]. Results conclusively demonstrate that MIGT has improved the accuracy by 9% to 97%, while the corresponding precision, recall, and F1-scores are all greater than 97%. This represents a notable performance improvement in terms of both the false positive and negative rates.

To verify this improvement, the three species with the highest misclassification rate in [15] were selected for further critical analysis. They are respectively Vernonia amygdalina (species #1), Dioscorea persimilis (species #2), and Stachytarpheta jamaicensis (species #3). Tables III to V respectively display the corresponding precision, recall, and F1 values for these three species. Results again corroborate that by incorporating MIGT into the CNN classification process, an improvement of over 50% in precision, 81% in recall and 76% on F1-scores respectively is achieved. This is most prominent for species #1 which had the lowest classification rate in [15], though commensurate enhancements are also evident in species #2 and #3, thus underscoring the rationale for the MIGT algorithm of applying MI to increase the intra-class coverage of samples for the training, validation, and testing sets used in the CNN model for classification.
TABLE II
CLASSIFICATION PERFORMANCE

<table>
<thead>
<tr>
<th>Method</th>
<th>Accuracy</th>
<th>Precision</th>
<th>Recall</th>
<th>F1-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Randomly selection [15]</td>
<td>88.26 %</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>MIGT</td>
<td>97.31 %</td>
<td>97.33 %</td>
<td>97.00 %</td>
<td>97.00 %</td>
</tr>
</tbody>
</table>

TABLE III
PRECISION IMPROVEMENT IN THREE SPECIES

<table>
<thead>
<tr>
<th>Method</th>
<th>species #1</th>
<th>species #2</th>
<th>species #3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Randomly selection [15]</td>
<td>0.30</td>
<td>0.63</td>
<td>0.57</td>
</tr>
<tr>
<td>MIGT</td>
<td>0.94</td>
<td>0.95</td>
<td>0.95</td>
</tr>
</tbody>
</table>

TABLE IV
RECALL IMPROVEMENT IN THREE SPECIES

<table>
<thead>
<tr>
<th>Method</th>
<th>species #1</th>
<th>species #2</th>
<th>species #3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Randomly selection [15]</td>
<td>0.07</td>
<td>0.90</td>
<td>0.93</td>
</tr>
<tr>
<td>MIGT</td>
<td>0.82</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

TABLE V
F1-SCORE IMPROVEMENT IN THREE SPECIES

<table>
<thead>
<tr>
<th>Method</th>
<th>species #1</th>
<th>species #2</th>
<th>species #3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Randomly selection [15]</td>
<td>0.12</td>
<td>0.74</td>
<td>0.70</td>
</tr>
<tr>
<td>MIGT</td>
<td>0.88</td>
<td>0.98</td>
<td>0.98</td>
</tr>
</tbody>
</table>

V. CONCLUSION

Despite the increasing availability of various public apps, accurate identification of medicinal plant species in their natural environments remains a significant computer vision challenge. Most existing image datasets do not authentically reflect the wide differences in plant species’ shape, colour and size or effectively cover the range of intra-class variations and features. While CNN models are a powerful tool for classifying images, inadequate training set coverage can lead to poor classification performance. This paper has presented a novel mutual information guided training (MIGT) algorithm, which selects samples that are more representative of the intra-class variations for each species in the training, testing and validation sets used by the CNN model for enhanced discrimination of subtle features and variations. Using the medicinal plant VNPlants-200 dataset, a critical comparative evaluation has confirmed the MIGT algorithm improved the average classification accuracy to more than 97%, with crucially, the precision, recall, and F1-scores also consistently being higher than 97%, so demonstrating its reliability and robustness. The MIGT algorithm is also a scalable solution for medicinal plant classification in larger environments. Future work includes developing MI refinements in guiding the training sample selection process and generalising the solution by rigorously evaluating the MIGT performance on other plant datasets.

REFERENCES


