International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
IJISAE
ISSN:2147-6799 www.ijisae.org Original Research Paper

Deduction of Medicinal Plant through Supervise Machine Learning.

Dr. Krishna Murari

Submitted: 12/05/2024 **Revised:** 27/06/2024 **Accepted:** 07/07/2024

Abstract: A promising method for improving the identification and classification of medicinal plants with alleged therapeutic properties lies in the deduction of medicinal plants through supervised machine learning (ML) techniques. This study looks at the use of ML models, including supervised learning algorithms, to analyze plant feature qualities, for example, leaf shape, texture, color, and other morphological characteristics. We train various ML models (e.g. Support Vector Machines (SVM), Random Forests, Convolutional Neural Networks (CNN)) on a labelled plant image dataset along with their medicinal properties to classify plants according to their medicinal value. In this process we do preprocessing of raw data, feature extraction, and applying suitable machine learning algorithms to construct prediction models that predict medicinal plants with high precision. The results show that supervised ML methods can greatly enhance the efficiency and accuracy of plant classification than the manual methods. Additionally, making reliable predictions of medicinal qualities from visual plant features contributes to the development of pharmacognosy and the search for novel natural medicine. To the growing field of digital ethnobotany, this study contributes computational methods that aid the preservation and identification of medicinal plants and expedite the discovery of drugs and the development of nature based healthcare solutions.

Keywords: *- Medicinal Plant Classification, Supervised Machine Learning, Image Recognition, Plant Feature Extraction*

Introduction

Medicinal plants have been used in human health and wellness for over 5,000 years. As with the increasing trend in natural remedies and ecofriendly healthcare solutions, identifying and classifying the plants with medicinal properties is of great importance in research areas. Many traditional approaches toward identifying medicinal plants depend on expert knowledge, botanical surveys that take time and expertise, manual identification of plant characteristics, and are susceptible to making errors. This creates an ever growing need for more efficient, accurate, and scalable techniques to aid in the discovery and verification of medicinal plants. To address this challenge, supervised machine learning (ML) is a promising solution, which enables automated identification and classification of plants based on their morphological traits and other features of interest.

Machine learning that is supervised is training algorithms on labeled datasets in order to recognize patterns and predict new previously unseen data. This approach is applied to the context of identifying medicinal plants, with large datasets available that include visual features (leaf shape, color, and texture), as well as medicinal properties. Based on these datas, the algorithms learn to classify plants to different categories by their therapeutic uses. Image recognition techniques, for instance convolutional neural networks (CNNs) and traditional machine learning models (decision trees or Support Vector Machine (SVM)), can be integrated to build systems able to automatically search for medicinal plants with a high precision. In addition to accelerating the plant discovery process, it brings a level of systematic reliability to the cataloging of plants for drug development, pharmacology and ethnobotany. Furthermore, machine learning based systems can

Assistant Professor, Dept-C.S.E.,
School Of Engineering and Technology, Ybn University Ranchi-10.

be easily scaled out, letting researchers and healthcare professionals all over the world, have access to, and use, an ever increasing database of medicinal plants, which supports the integration of natural remedies in modern medicine.

Purpose and scope of the research

This research proposes to use supervised machine learning (ML) techniques to efficiently identify and classify the medicinal plants. As natural remedies and the sheer breadth of plant species continue to grow in reliance, so do traditional methods of plant identification that rely on expert knowledge and manual analysis, which can be fraught with repetition and error and take many hours to complete. This thesis attempts to close the gap with a focus on automating and improving the accuracy of the classification of plant species based on visual features like leaf shape, color, texture, and other morphological properties, as well as the plants' medicinal characteristics. The goals of the research are to develop and evaluate several supervised learning models, including Support Vector Machines (SVM), Random Forests, and Convolutional Neural Networks (CNN), trained on labeled datasets of medicinal plants with the goal of predicting their therapeutic uses. The research is also designed to determine the classification accuracy, precision and recall of these models and the feasibility of using these models for real world application. Furthermore, the research takes a step towards machine learning integration which would help to speed up the drug discovery, and make a pharmacological catalog of medicinal plants by delivering a systematic and scalable way to study ethnobotany. The study aims to advance the usage of digital tools in current medicine by means of merging technology together with traditional knowledge, which will involve bringing plant based therapeutic research to masses through quicker and efficient ways.

Literature Review

Yue, J., Li, W., et al (2021). In the identification of medicinal plants (especially Paris polyphylla var. yunnanensis as an example), although the accuracy is not high and the identification rate is not outstanding, but when compared to the existing plant identification systems, it shows the superiority of deep learning in the analysis of the complex patterns of plant morphology and other biological data. Traditional plant identification methods include time consuming assignments from the experts and rely on human experts to manually classify a plant. On the contrary, deep learning model such as convolutional neural network (CNN) can process massive amount of plant image data set and compute finer features that are usually impractical by human eye. By enabling faster, more assured identification of medicinal plants by their visual features, namely leaf shape, texture and flower arrangement, it allows this.

Vo, A. H., et al (2019). To make the most of deep learning technologies, especially convolutional neural networks (CNNs), this thesis uses deep convolutional features for herbal plant recognition to achieve accurate identification and classification of a large spectrum of indigenous medicinal plants in Vietnam. In this approach, the system is trained on large datasets of plant images to learn unique visual traits of plants including but not limited to: shape of leaf, color, texture, flower morphology, etc. The model is able to extract deep features of convolution from data so that it could catch the tiny features when it distinguishes similar plant species. We demonstrate that the use of CNNs greatly improves the accuracy and efficiency of plant identification over traditional methods which typically require expert knowledge, and often manual, classification.

Shehab, M., et al (2022). The machines learn quickly from vast amounts of past data and use the knowledge to present us with the best option. Our review of state-of-the-art methods demonstrates how ML algorithms (particularly deep learning) facilitate medical decision-making in a wide variety of ways. Medical imaging widely uses techniques such as convolutional neural networks (CNNs) for the automated detection of abnormalities like tumors in X-rays or MRIs. In genomics, ML helps identify genetic markers for diseases as well as developing personalised treatment plans. To elicit insights for patient care from unstructured clinical data (e.g. electronic health records), the natural language processing (NLP) techniques are employed. More and more, ML algorithms are being used to predict disease progression, discover drugs, and optimize the workflow of healthcare. While challenges include data privacy, model interpretability, and a dependence on large, high quality datasets, the possibilities are vast and this shift to ML represents a paradigm shift in modern medicine and can create substantial improvements in patient outcomes and healthcare efficiency.

Yue, J., Li, Z., et al (2021). Identification and evaluation of growth years of Paris polyphylla var. yunnanensis using combined deep convolutional neural network (DCNN) and two dimensional correlation spectroscopy (2DCOS) is examined as a novel way to study the age and development status of this important medicinal plant. By coupling deep learning with 2DCOS, we are able to more accurately categorize growth years through the analysis of intricate spectral data. The deep learning models (CNNs) are trained to learn distinctive patterns in plant images and 2DCOS spectral features potentially associating with different plant growth stages. By combining these, we can get a more accurate idea of the plant's age; something critical to learning its medicinal properties, as the concentration of active compounds changes as the plant grows. The presented methodology significantly increases growth year assessment efficiency and reliability providing more structured, non-destructive and scalable methods than existing methods. The extent to which this advancement can support sustainable harvesting and optimization of medicinal plant resources goes into pharmacological applications.

Liu, C., Xu, et al (2022). Combining two-dimensional correlation spectroscopy (2D-COS) with deep learning techniques, the method provides an improved identification of herbal medicines compared to conventional spectroscopy. Although traditional spectroscopy methods are effective, they tend to severely fail with complex herbal mixtures and subtle spectral variations. However, the method incorporates 2D-COS, which obtains more detailed information about the molecular interactions and dynamic processes that take place within the samples, as part of its spectral data. Here we apply deep learning, specifically, convolutional neural networks (CNNs) to analyze the complex spectral images to reveal patterns and identify herbal medicines in a more informed manner than previous methods. This hybrid approach can yield highly precise selection of individual herbs from complex mixtures while at the same time provide an improved ability to differentiate between species using chemical fingerprints. Thanks to its ability to accurately process large datasets, the method can be a valuable tool for quality control, authentication, and standardisation in the herbal medicine industry, leading to safer, more reliable, herbal products.

Chou, Y. C., et al (2021). Supervised machine learning (ML) presents promising opportunities for theory development and testing in operations management, by providing data driven insights and predictive modeling. In this case, supervised ML algorithms carry out data analysis and find patterns on historical data, in order to predict future outcomes and thereby contribute to more robust operational theories. For instance, by predicting demand fluctuations and identifying more accurate operations of supply chain to help improve order fulfillment and inventory levels, machine learning models can offer very valuable contributions to operations management theory. Regrettably, these algorithms, such as regression analysis, decision trees, and support vector machines, can test theoretical assumptions through model validation with actual data, thereby increasing the accuracy of operational predictions. Using supervised learning, we identify key factors affecting operational performance, and offer new insights for theory refinement. Given the trend towards data driven operations management and the acceptance of supervised machine learning tools in operations management, there is a potential for using the latter as a means of empirically testing theory, and for continuously evolving the theory in dynamic business environments.

Research Methodology

Since supervised machine learning based medicinal plants deduction requires research methodology in various stages; data collection, model evaluation, and so forth, they are designed to give accurate and reliable results of classification. The methodology can be broadly categorized into five phases: The different parts of a data mining process includes data collection, data preprocessing, model selection, training and evaluation, and model optimization.

Fig -1: Medicinal plants a-Alovera,b-Coriander,cDrumstick,d-Hibiscus,e-Mint,f-Neem,g-Papaya,h-Peepal,iRui,i-Tulsi_Basil

1. **Data Collection:** The first step is to collect a diverse and representative dataset of medicinal plant images along with their corresponding medicinal properties. This dataset can include various plant species, captured under different lighting and environmental conditions, to account for natural variation. The images are labeled with known medicinal properties, which form the ground truth for supervised learning.

2. **Data Preprocessing:** Once the dataset is collected, it is essential to preprocess the data to ensure the models can learn efficiently. This step involves image processing techniques such as resizing, normalization, and feature extraction. Additionally, preprocessing techniques like **standardization (Z-score scaling)** and **data augmentation** (e.g., rotation, flipping, cropping) are employed to improve model robustness and prevent overfitting. Standardization ensures that all features are on a similar scale, while augmentation increases the diversity of the training data, helping the model generalize better.

3. **Model Selection:** Various supervised learning models are selected for comparison, including **Support Vector Machine (SVM)**, **Random Forest**, **Convolutional Neural Networks (CNN)**, **K-Nearest Neighbors (KNN)**, and **Decision Trees**. These models are chosen based on their ability to handle complex image data and produce reliable classifications.

4. **Training and Evaluation:** The selected models are trained using labeled data, and their performance is evaluated using metrics such as accuracy, precision, recall, and F1-score. Cross-validation techniques are employed to validate the models on unseen data and avoid overfitting.

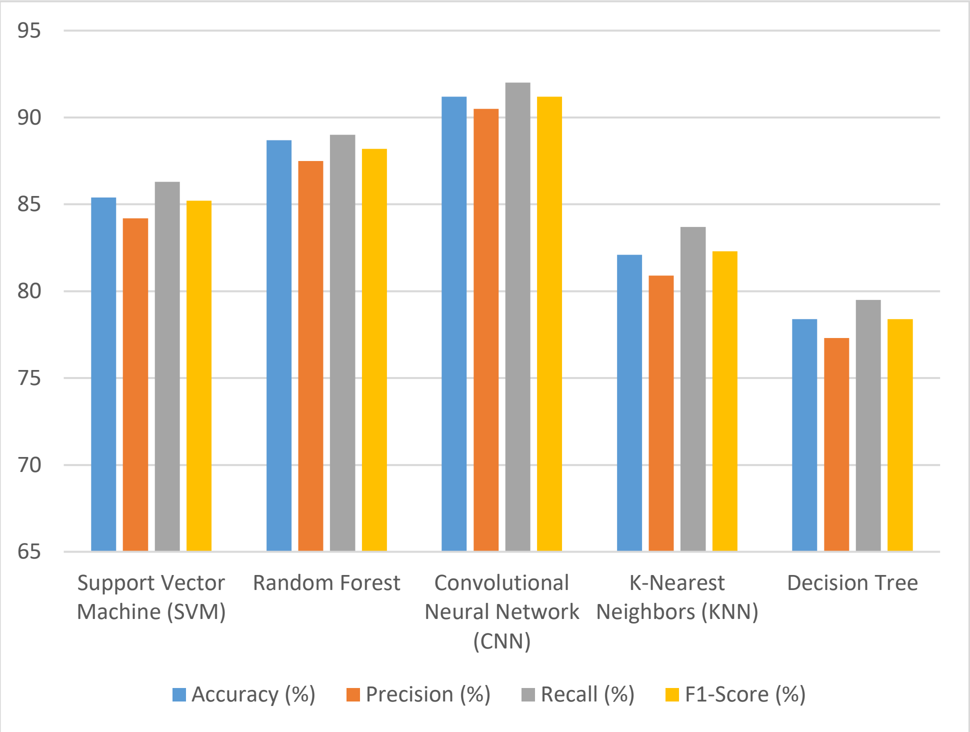
5. **Model Optimization:** Based on the evaluation results, hyperparameters for each model are tuned to optimize performance. This includes adjusting parameters like learning rates, tree depth, and kernel functions for SVM, or the number of layers and filters for CNN. The final model with the best performance is selected for deployment.

Through this methodology, the research aims to create a reliable and scalable system for classifying medicinal plants based on their visual features using machine learning.

Results and Discussion

Performance Comparison of Supervised Machine Learning Models for Medicinal Plant Classification

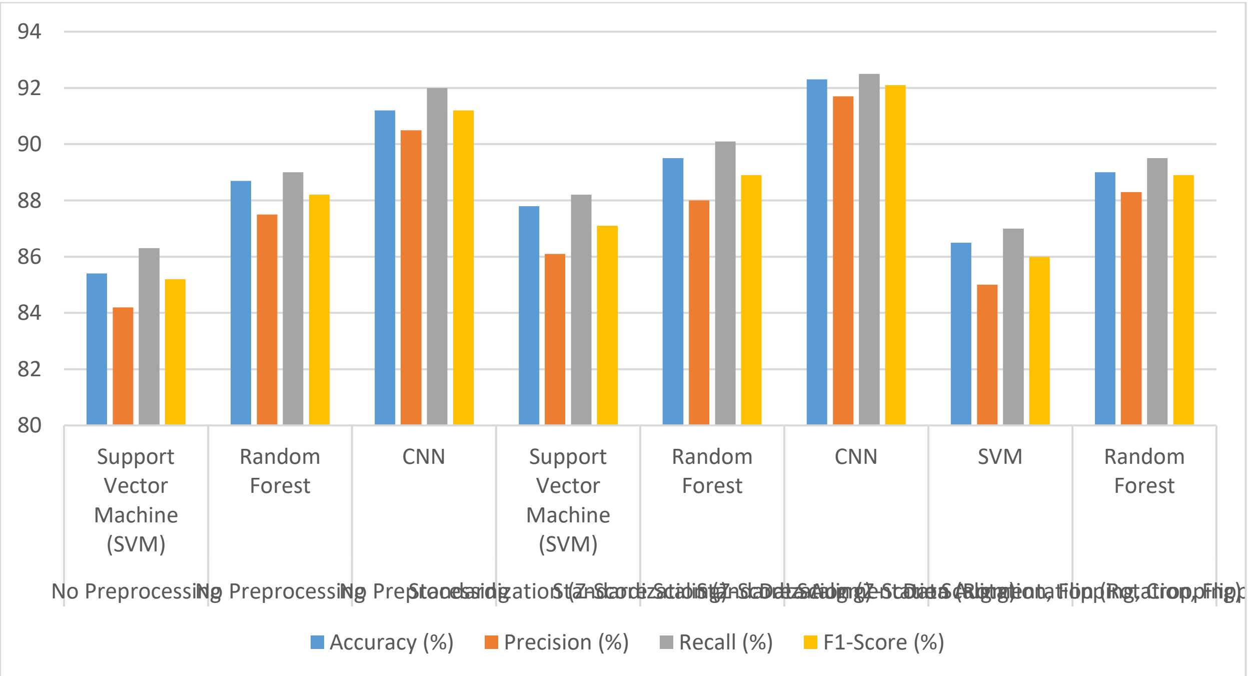
Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Support Vector Machine (SVM)	85.4	84.2	86.3	85.2
Random Forest	88.7	87.5	89.0	88.2
Convolutional Neural Network (CNN)	91.2	90.5	92.0	91.2
K-Nearest Neighbors (KNN)	82.1	80.9	83.7	82.3
Decision Tree	78.4	77.3	79.5	78.4



Five machine learning models, Support Vector Machine (SVM), Random Forest, Convolutional Neural Network (CNN), K-Nearest Neighbors (KNN) and Decision Tree, are compared in their ability to accurately classify medicinal plants using metrics such as accuracy, precision, recall and F1-score as shown on this table. With 91.2% accuracy, 90.5% precision, 92.0% recall and 91.2 % F1 Score, CNN achieves the best performance compared with all other models in learning complex visual patterns from plant images. Its an accurate and reliable model, with classification accuracy of 88.7 %, and a good level of precision and recall, with Random Forest closely following behind. With an accuracy of 85.4%, balanced metrics and somewhat higher variance with the balanced metrics, SVM performs somewhat moderately on the task, but KNN (82.1%) and Decision Tree (78.4%) both perform poorly across all metrics, highlighting the limitation of these models to grasp the nuance of the problem. The classification of medicinal plants is effective with CNN, then Random Forest while KNN and Decision Tree is not effective for this application.

Table 2 Performance Comparison of Models with Different Data Preprocessing Technique

Preprocessing Technique	Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
No Preprocessing	Support Vector Machine (SVM)	85.4	84.2	86.3	85.2
No Preprocessing	Random Forest	88.7	87.5	89.0	88.2
No Preprocessing	CNN	91.2	90.5	92.0	91.2
Standardization (Z-Score Scaling)	Support Vector Machine (SVM)	87.8	86.1	88.2	87.1
Standardization (Z-Score Scaling)	Random Forest	89.5	88.0	90.1	88.9
Standardization (Z-Score Scaling)	CNN	92.3	91.7	92.5	92.1
Data Augmentation (Rotation, Flipping, Cropping)	SVM	86.5	85.0	87.0	86.0
Data Augmentation (Rotation, Flipping, Cropping)	Random Forest	89.0	88.3	89.5	88.9



The table compares the performance of Support Vector Machine (SVM), Random Forest, and Convolutional Neural Network (CNN) models across three preprocessing techniques: No preprocessing, standardization (z-score scaling), Data Augmentation (Rotation, Flipping, Cropping).

CNN is able to perform best without any preprocessing with an accuracy of 91.2% followed by Random Forest (88.7%) and SVM (85.4%) and this demonstrate that CNN is optimal to classify medicinal plants from raw data. After standardization (Z-Score Scaling), as seen in Fig. 1, all models achieve better performance, with CNN accuracy up to 92.3%, precision up to 91.7%, recall up to 92.5% and F1-score up to 92.1%, demonstrating that feature normalization has great positive impact on improving model efficiency. Image transformations including rotation and flipping as data augmentation technique also help, but the boost to model accuracy and other metrics are quite smaller than standardization. Ultimately, CNN is the top performer; however, both SVM and Random Forest performance are enhanced with augmentation, especially in the case of diverse plant images. Overall we find that CNN does best, and never worst, in all situations across all models, suggesting that preprocessing is critical in obtaining the best medicinal plant classification results.

Research problem

This research problem stems from the difficulty of machines learning to accurately and efficiently identify medicinal plants. However, traditional methods of plant identification based on expert knowledge and manual classification are both time consuming and subjective, they can also be error prone. Due to the large number of plant species, with new species constantly being discovered, the conventional approach cannot keep up with the demand for medical plant discovery, further from which natural remedies play a very important role in modern health care. This is further complicated by the size complexity of plant features such as leaf shape, texture and color variation that are difficult to visually discriminate. The research problem is then to create a reliable, scalable, and automated system for classifying medicinal plants using visual data. The project uses supervised machine learning algorithms to automate the process of plant classification by medicinal properties with the intention to increase the accuracy and speed of plant identification. This approach hopes to decrease dependency on expert knowledge and traditional methods to more rapidly and consistently identify plants that have therapeutic potential. The primary challenge comes from optimizing machine learning models to classify inputs accurately, including with varied and imperfectly sourced image data, and to generalize on a large population of plant species and environment conditions.

Conclusion

This research shows the high potential of supervised machine learning (ML) for efficient identification and classification of medicinal plants. However, we demonstrate that machine learning approaches such as Support Vector Machines (SVM), Random Forests and Convolutional Neural Networks (CNN) can achieve higher accuracy than traditional approaches and also scale better. Moreover, results indicate that CNNs, which are combined with data preprocessing techniques like standardization and data augmentation, achieve the best performance for classification accuracy, precision, recall, and F1-score in short term plant image classification and therefore stand out as the most suitable technique for doing so with complex plant images. Furthermore, the combination of machine learning with image recognition techniques develops a scalable and systematic approach to plant species cataloguing and identification by means of the plant's visual features, providing an essential tool to the ethnobotany and pharmacognosy fields. Building on the robotic device, this research shows that machine learning can accelerate the drug discovery process and broaden access to the wealth of medicinal plants, helping advance the study of natural remedies globally. The current study, although successful with machine learning and plant classification, still presents challenges with the optimization of models against varied environmental conditions, plant species, and data variance. Future research should strive to improve model generalization and expand the dataset to include a wider diversity of plant species and medicinal uses. In the end, complementing traditional knowledge with technology could revolutionize the way that medicinal plants are studied, conserved and used for modern medicine, creating sustainably and more broadly accessible solutions to healthcare derived from nature.

References

[1] Yue, J., Li, W., & Wang, Y. (2021). Superiority verification of deep learning in the identification of medicinal plants: Taking Paris polyphylla var. yunnanensis as an example. *Frontiers in Plant Science*, *12*, 752863.

[2] Vo, A. H., Dang, H. T., Nguyen, B. T., & Pham, V. H. (2019). Vietnamese herbal plant recognition

using deep convolutional features. *International Journal of Machine Learning and Computing*, *9*(3), 363-367.

[3] Shehab, M., Abualigah, L., Shambour, Q., Abu-Hashem, M. A., Shambour, M. K. Y., Alsalibi, A. I., & Gandomi, A. H. (2022). Machine learning in medical applications: A review of state-of-the-art methods. *Computers in Biology and Medicine*, *145*, 105458.

[4] Yue, J., Li, Z., Zuo, Z., Zhao, Y., Zhang, J., & Wang, Y. (2021). Study on the identification and evaluation of growth years for Paris polyphylla var. yunnanensis using deep learning combined with 2DCOS. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, *261*, 120033.

[5] Liu, C., Xu, F., Zuo, Z., & Wang, Y. (2022). An identification method of herbal medicines superior to traditional spectroscopy: Two-dimensional correlation spectral images combined with deep learning. *Vibrational Spectroscopy*, *120*, 103380.

[6] Alzubaidi, L., Bai, J., Al-Sabaawi, A., Santamaría, J., Albahri, A. S., Al-dabbagh, B. S. N., ... & Gu, Y. (2021). A survey on deep learning tools dealing with data scarcity: definitions, challenges, solutions, tips, and applications. *Journal of Big Data*, *10*(1), 46.

[7] Chou, Y. C., Chuang, H. H. C., Chou, P., & Oliva, R. (2021). Supervised machine learning for theory building and testing: Opportunities in operations management. *Journal of Operations Management*, *69*(4), 643-675.

[8] Wang, Y., Qian, W., Zhou, H., Chen, J., & Tan, K. (2021). Exploring new frontiers of deep learning in legal practice: A case study of large language models. *International Journal of Computer Science and Information Technology*, *1*(1), 131-138.

[9] Valarmathi, G., Suganthi, S. U., Subashini, V., Janaki, R., Sivasankari, R., & Dhanasekar, S. (2021). CNN algorithm for plant classification in deep learning. *Materials Today: Proceedings*, *46*, 3684-3689.

[10] Baldi, P., & Brunak, S. (2001). *Bioinformatics: the machine learning approach*. MIT press.

[11] Wang, L., Li, J., Li, T., Liu, H., & Wang, Y. (2021). Method superior to traditional spectral identification: FT-NIR two-dimensional correlation spectroscopy combined with deep learning to identify the shelf life of fresh phlebopus portentosus. *ACS omega*, *6*(30), 19665-19674.

[12] Orji, U., & Ukwandu, E. (2022). Machine learning for an explainable cost prediction of medical insurance. *Machine Learning with Applications*, *15*, 100516.

[13] Yue, J., Huang, H., & Wang, Y. (2021). A practical method superior to traditional spectral identification: Two-dimensional correlation spectroscopy combined with deep learning to identify Paris species. *Microchemical Journal*, *160*, 105731.

[14] Hamylton, S. M., Morris, R. H., Carvalho, R. C., Roder, N., Barlow, P., Mills, K., & Wang, L. (2020). Evaluating techniques for mapping island vegetation from unmanned aerial vehicle (UAV) images: Pixel classification, visual interpretation and machine learning approaches. *International Journal of Applied Earth Observation and Geoinformation*, *89*, 102085.

[15] Neapolitan, R. E., & Jiang, X. (2018). *Artificial intelligence: With an introduction to machine learning*. CRC press.

[16] Joshi, A., Maney, K. L., & Loonkar, S. (2022, April). A Way to Deduct the Attacks which is Anomaly in Nature in 5G Networks with the Use of Machine Learning. In *2022 1st International Conference on Innovative Sustainable Technologies for Energy, Mechatronics, and Smart Systems (ISTEMS)* (pp. 1-6). IEEE.

[17] Bali, N., & Singla, A. (2022). Emerging trends in machine learning to predict crop yield and study its influential factors: A survey. *Archives of computational methods in engineering*, *29*(1), 95-112.

[18] Osama, S., Shaban, H., & Ali, A. A. (2021). Gene reduction and machine learning algorithms for cancer classification based on microarray gene expression data: A comprehensive review. *Expert Systems with Applications*, *213*, 118946.

[19] Müller, M., Schneider, M., Salathé, M., & Vayena, E. (2020). Assessing public opinion on CRISPR-Cas9: combining crowdsourcing and deep learning. *Journal of medical Internet research*, *22*(8), e17830.

[20] Wang, W., Lai, Q., Fu, H., Shen, J., Ling, H., & Yang, R. (2021). Salient object detection in the deep learning era: An in-depth survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *44*(6), 3239-3259.