Proceedings of the 4th International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/49/20241423

Research on improving accuracy and efficiency of animal data collection and classification using machine learning

Qinlong Yang

University of Science and Technology Beijing, No. 30 Xueyuan Road, Haidian District, Beijing, China

42024185@xs.ustb.edu.cn

Abstract. With the continuous expansion of machine learning algorithms in various application domains, the application value of new algorithms, such as Support Vector Machines and Convolutional Neural Networks in data classification, has garnered increasing attention. This paper takes machine learning algorithms as the research entry point, explores the concept of machine learning, and delves into its application value in data classification. This paper, starting with an overview of machine learning algorithms, analyzes the supervised and unsupervised learning problems in machine learning, focusing on the applications of Convolutional Neural Networks, Support Vector Machine models, and logistic regression algorithms in data classification. This study emphasizes designing and implementing a machine learning-based image classification system. Through an in-depth exploration of the application of machine learning algorithms in data classification, a fully functional system is constructed, encompassing multiple modules, including machine vision and software development. This system accurately classifies and recognizes images, providing practical tools and technical support for image processing and analysis. In this study, the goal of achieving good image classification is realized through research and the application of machine learning algorithms. By designing and implementing a machine learning-based image classification system, the accuracy and efficiency of classification in handling massive data are improved. This system also demonstrates wide-ranging prospects in software development and machine vision, among other fields.

Keywords: Machine learning, software development, functional modules, machine vision.

1. Introduction

Machine learning algorithms, such as Support Vector Machines (SVM) and Convolutional Neural Networks (CNN), are gaining attention for data classification as the internet and information industry grow exponentially. Handling this vast data requires advanced technology, and machine learning offers the solution. This study focuses on machine learning's application in data classification, particularly image classification. It aims to improve classification accuracy and efficiency by developing a machine learning-based image classification system, with potential applications in software development and machine vision.

This paper aims to design and implement a machine learning-based image classification system. It explores the application of machine learning algorithms, including CNNs, SVM models, and logistic regression, to construct an accurate image classification system. The goal is to enhance image data

DOI: 10.54254/2755-2721/49/20241423

processing accuracy and efficiency while contributing to software development and machine vision technology.

This research is significant for designing and implementing an image classification system using machine learning. It addresses the challenge of handling the exponential growth of network information in the information industry. The developed system enables accurate image classification and recognition, serving as a valuable tool for image processing and analysis. It holds practical value in software development and machine vision and contributes to advancing image classification and recognition technology.

2. Literature review

The use of machine learning and radiomic features in medical imaging for disease diagnosis and prediction has gained considerable attention [1]. This review summarizes studies across various medical applications, such as pituitary prolactinoma diagnosis, renal cancer subtype classification, rectal cancer staging, white blood cell classification, object detection in images, assembly quality recognition, nematode identification, ship image anomaly detection, underground drainage pipe defect classification, and thyroid nodule ultrasound image recognition [2].

Pituitary prolactinoma is challenging to diagnose accurately due to image similarities between different tumor types. Researchers have explored radiomic features and machine learning models, like SVM and deep learning, to improve diagnosis accuracy and tumor localization. Results show significant potential for diagnosing and localizing pituitary prolactinomas, aiding surgical planning [3].

Classifying renal cancer subtypes is essential for personalized treatment planning. Utilizing radiomic features from enhanced CT scans, machine learning models achieved high accuracy in distinguishing subtypes, offering a non-invasive alternative to histological examination.

Automated T staging of rectal cancer using T2WI and RS-EPI DWI radiomic features improved preoperative assessments. Machine learning models accurately predicted tumor staging, enhancing surgical planning and treatment decisions.

Machine learning-based white blood cell classification automated a traditionally manual process. The model effectively categorized white blood cell types, improving efficiency and aiding in disease diagnosis and monitoring [4].

An approach combining color and depth information with sample selection improved object detection in RGBD images. The method enhanced accuracy while reducing false positives, providing an effective solution for object detection.

Machine learning-based recognition of assembly quality in manufacturing improved quality control efficiency [5]. The model detected quality issues, aiding in reducing non-conforming product rates.

Flow-based nematode identification on microfluidic chips combined with machine learning provided high-throughput classification and analysis. The system rapidly categorized nematodes, benefiting biological research.

Machine learning enhanced the detection of anomalous targets in multispectral ship images, improving ocean monitoring and ship recognition accuracy [6].

Machine learning effectively classified underground drainage pipe defect images, enhancing detection and classification efficiency for urban infrastructure maintenance [7].

Radiomic features and machine learning aided in thyroid malignant nodule ultrasound image recognition and Traditional Chinese Medicine tongue diagnosis classification. These methods hold promise for personalized medical decisions and improved patient outcomes [8].

3. Animal image classification experiment and results

3.1. Dataset description

3.1.1. Dataset source. This paper sourced This study dataset from Kaggle, which offers a comprehensive collection of annotated animal images, including endangered species [9]. These images

DOI: 10.54254/2755-2721/49/20241423

were obtained from wildlife monitoring sites in national parks, nature reserves, and wildlife habitats. this paper also utilized resources like the Biodiversity Image Library and Natural History Museum databases, aggregating biodiversity images from around the world. This study dataset covers a wide range of animal species, from large mammals to rare birds, and spans various ecosystems and geographic locations [2]. This diversity enhances the performance and applicability of this study animal image classification system, benefiting ecological research and wildlife conservation.

3.1.2. Dataset size and characteristics. This study dataset includes images representing different animal phyla, classes, and species, encompassing large mammals (e.g., lions, elephants), aquatic animals (e.g., dolphins, sea turtles), terrestrial birds (e.g., eagles, peacocks), reptiles (e.g., snakes, crocodiles), and insects (e.g., butterflies, ants). These images showcase a variety of ecosystems, including forests, grasslands, deserts, aquatic environments, and high-altitude regions. However, the dataset exhibits class imbalance, with some species having more images than others. This imbalance poses challenges in classification and requires special handling. All images are accompanied by accurate annotations, including species and behavioral characteristics, facilitating supervised learning tasks.

3.2. Machine learning algorithm selection

3.2.1. Convolutional neural network (CNN). CNNs are ideal for image processing, as they can automatically learn features from images without manual feature extraction. They consist of multiple layers for hierarchical feature extraction and can be fine-tuned for specific tasks, making them well-suited for this study's diverse animal image dataset.

3.2.2. Support vector machine (SVM). SVM is a classical classification algorithm known for its effectiveness in image classification. It can handle multiclass problems, and high-dimensional data, and exhibits good generalization to unseen data, essential for this study task's ecological diversity.

3.2.3. Logistic regression algorithm. Logistic regression is a simple yet interpretable classifier suitable for baseline models and initial attempts. It provides probability estimates and insight into feature importance, aiding this study's understanding of classification decisions.

3.3. Experimental design and methods

3.3.1. Train-Test split. This paper divided the dataset into training, validation, and test sets, ensuring a representative distribution of animal categories in each. This adhered to cross-validation principles to ensure reliable results.

3.3.2. Data augmentation strategy. Data Augmentation: To improve model generalization, this paper applied data augmentation techniques. This involved random rotations, flips, cropping, brightness adjustments, and color distortions to simulate varying image conditions.

3.3.3. Model training and Fine-Tuning. CNN models were initialized with pre-trained weights and fine-tuned on this study dataset. SVM and logistic regression models used feature vectors extracted by the CNN as inputs and were trained accordingly.

3.3.4. Performance evaluation metrics. This paper used classification accuracy, precision, recall, F1 score, ROC curves, and AUC values to evaluate model performance comprehensively.

DOI: 10.54254/2755-2721/49/20241423

4. Image classification system design and implementation

4.1. System functional module design

4.1.1. Data preprocessing module. Data preprocessing is essential for data quality. It involves removing noise, errors, and outliers in collected animal images [1]. Standardizing image sizes ensures uniform processing and eliminates biases. Data augmentation introduces transformations like rotation and cropping, increasing training data diversity, improving model robustness, and reducing overfitting.

4.1.2. Feature extraction module. CNNs automatically extract features from images. CNN layers detect low-level features and progressively combine them into high-level semantic representations [7]. CNN's ability to learn features without manual design makes it adaptable to diverse and complex animal image datasets. It transforms raw images into high-level feature representations for classification.

4.1.3. Classifier model selection. SVM and Logistic Regression are chosen for their generalization ability. SVM separates data points with an optimal hyperplane, handling high-dimensional data and nonlinear relationships. Logistic regression classifies data and maintains good performance with new, unseen data. Both models use feature representations from CNN for image classification.

4.1.4. User Interface Module. The user interface prioritizes user-friendliness, ensuring effortless interaction regardless of technical proficiency [10]. It offers clear layouts, intuitive image upload, and guidance for easy system operation. Users can upload images, and the system immediately processes and classifies them, providing quick results in a user-friendly visual format. The interface is cross-platform compatible for widespread accessibility.

4.2. Computer vision module

4.2.1. CNN architecture. This study CNN model, with multiple convolutional and pooling layers, captures abstract image features automatically. It scans images using convolutional kernels to identify edges, textures, and shapes. Pooling layers reduce complexity while preserving critical feature information. Pre-trained weights, often from models trained on datasets like ImageNet, are used to enhance the model's generalization. Optimizers and loss functions are chosen to minimize training loss.

4.2.2. Model training. The training process involves the training, validation, and testing sets. Backpropagation optimizes model weights through iterations, gradually improving performance. Dropout and regularization techniques prevent overfitting [8]. Model training ensures the model captures animal image features and maps them to categories, with careful tuning and validation for generalization.

4.2.3. Model Fine-Tuning. This paper fine-tuned a pre-trained model on this study animal image dataset to adapt it to the specific task. Fine-tuning involves adjusting weights, architecture, and hyperparameters. Regularization techniques and hyperparameter tuning optimize model performance for accurate classification.

4.3. Software development module

4.3.1. Coding implementation. Python, TensorFlow, and Scikit-Learn are used for coding. TensorFlow provides deep learning tools, while Scikit-Learn is used for traditional machine learning. Web frameworks like Django or Flask help build user interfaces, and database systems like MySQL or SQLite manage data. Cloud platforms like AWS or Azure ensure system availability and scalability.

DOI: 10.54254/2755-2721/49/20241423

4.3.2. Integration testing. Integration testing validates module interactions, data flow, and system performance [4]. Exception scenarios are simulated to test error handling and recovery mechanisms. Regression testing prevents code changes from breaking existing functionality, ensuring system stability.

4.3.3. User interface design. The user interface simplifies image uploading through drag-and-drop or button selection. Real-time feedback informs users of image processing, reducing waiting anxiety. Classification results, including labels and confidence scores, are displayed clearly. Error messages guide users in case of issues. The interface is responsive and works seamlessly on various devices.

5. System performance evaluation

5.1. Dataset selection

This study dataset is sourced from various geographical regions worldwide, showcasing animals from different ecosystems such as forests, grasslands, and water bodies. This diversity allows us to evaluate animal classification across diverse environments.

The dataset includes a wide range of animal species, from large mammals like lions to small mammals like squirrels. It encompasses birds, insects, and aquatic animals, offering a comprehensive testbed for this study classification system's performance.

This study dataset covers various ecological environments, including forests, grasslands, aquatic ecosystems, and high-altitude regions. This diversity in natural backgrounds ensures this study system's adaptability to different ecological conditions.

Each image in the dataset is meticulously annotated with accurate species information and possible behavioral characteristics. These annotations are provided by experts and greatly enhance the credibility and utility of this study classification system.

5.2. Evaluation metrics

To comprehensively assess the performance of this study-designed and implemented animal image classification system, this paper has selected multiple evaluation metrics that help us gain a deeper understanding of the system's performance and provide multidimensional performance analysis. Accuracy is a core evaluation metric, representing the model's overall classification correctness on the entire test dataset, i.e., the proportion of images correctly classified. Accuracy is typically used to measure the model's overall performance. Precision represents the proportion of actual positive samples among all samples predicted as positive by the model. This metric measures the accuracy of the model in predicting positive samples and helps identify cases of misclassification. The recall represents the proportion of actual positive samples that the model correctly predicted as positive. This metric emphasizes the model's ability to capture positive samples and helps identify missed positive samples. The F1 score is a composite metric of precision and recall and is used to balance the model's accuracy and comprehensiveness. It is especially useful for handling class-imbalanced situations and aids in evaluating the performance balance between positive and negative classes. The Receiver Operating Characteristic Curve (ROC curve) is used to assess the binary classification performance of the model. It displays the trade-off between the True Positive Rate and the False Positive Rate at different thresholds. The Area Under the ROC Curve (AUC) is the area under the ROC curve and is commonly used to measure the model's classification performance. These metrics are particularly suitable for evaluating models like Support Vector Machines (SVM) and logistic regression.

5.3. Experimental results and analysis

5.3.1. Model performance comparison. In this study experiments, this paper evaluated the performance of multiple machine learning models, including Convolutional Neural Networks (CNNs), Support Vector Machines (SVMs), and logistic regression algorithms. This study's goal was to assess the

DOI: 10.54254/2755-2721/49/20241423

strengths and weaknesses of these models for animal image classification using various metrics. this paper considered accuracy, precision, recall, and the F1 score to comprehensively compare these models. Accuracy measures overall correctness, while precision assesses the accuracy of positive predictions. Recall focuses on capturing positive samples, and the F1 score balances precision and recall, especially in class-imbalanced scenarios.

*5.3.2. Handling class imbalanc*e. This study dataset exhibits class imbalance, with some animal categories having fewer samples. To address this, this paper applied weighting techniques during training. By adjusting sample weights, this paper ensured that the model paid more attention to minority classes, mitigating the impact of class imbalance. This paper used resampling strategies to balance class distributions, including oversampling and undersampling. Ensemble methods, such as ensemble learning and stacking, were employed to enhance performance further and manage class imbalance effectively.

5.3.3. Impact of ecological environments. This study dataset covers various ecological environments, each with unique challenges like varying lighting, weather conditions, and interfering factors. this paper assessed the model's adaptability to these real-world conditions. This research analyzed the model's performance under different lighting conditions, including cloudy, sunny, and rainy weather. Furthermore, this paper evaluated its ability to differentiate between different animal species, as well as its performance in the presence of environmental noise and interference.

6. Conclusion

This paper has provided a comprehensive overview of the design, implementation, and performance evaluation of an integrated animal image classification system. This study system leverages machine learning and deep learning technologies to automate the identification and categorization of various animal images. This paper has constructed a diverse and representative dataset encompassing distinct geographic regions, ecological environments, and animal species. This dataset is the foundation of this study system, ensuring its capability to address various challenges and scenarios. This paper has chosen multiple evaluation metrics, including accuracy, precision, recall, F1 score, and AUC value, to comprehensively assess the system's performance. These metrics have provided us with multi-dimensional insights into the performance of the classification system. By adopting a multi-model strategy, including CNNs and logistic regression models, this paper has extensively evaluated their performance in animal image classification. This approach has enabled us to select the optimal model as the system's core component.

Additionally, this paper has addressed the issue of class imbalance and implemented various strategies to improve the model's ability to handle imbalanced classes, enhancing the system's robustness and performance. This paper has also analyzed the model's performance under different ecological environments, lighting conditions, weather situations, animal species, and environmental noise to evaluate its adaptability in real-world scenarios. These analyses have contributed to This study's understanding of the model's limitations and potential areas for improvement.

Exploring more advanced deep learning architectures can enhance the system's classification accuracy. Utilizing pre-trained large-scale neural network models and fine-tuning them may further improve performance. Increasing the size and diversity of the dataset, including more geographic regions, ecological environments, and species, can enhance the system's generalization and adaptability. Exploring reinforcement learning techniques by simulating data under different environmental conditions is a promising research avenue. A critical research area is implementing real-time performance monitoring and enabling the system to learn and improve online to adapt to new animal species and environmental conditions. This would increase the practical applicability of the system.

Further optimizing the user interface for a more user-friendly experience, with consideration for mobile device applications, can meet a wide range of user needs. Applying the system to real-world scenarios, such as wildlife monitoring and ecological research, can validate its practical utility as it sees

DOI: 10.54254/2755-2721/49/20241423

broader adoption and ethical and privacy concerns become increasingly important. Future research should focus on handling and protecting sensitive ecological data, ensuring compliance with ethical principles.

References

[1] Kong X, Li W, Long YL, Meng M, Li YJ, Ma J. 2021 Diagnosis of Pituitary Prolactinoma Using Machine Learning Models Based on Radiomic Features *Chinese J. of Radiology.* **08**

[2] Yang G, He LY, Gao LG, He XW, Li DJ, Han C, Wang SY, Wu JY, Chen X 2019 Three-Class Prediction Model of Renal Cancer Subtypes Based on Radiomic Features from 3D Enhanced CT Images *J. of Func. Mater. and Dev.* **04**

[3] Wen DG, Hu SX, Li ZL, Deng XB, Tian C, Li X, Wang XR, Leng Q, Xia CC 2021 Value of an Automated Machine Learning Model Based on T2WI and RS-EPI DWI Radiomic Features in Predicting Preoperative T Staging of Rectal Cancer *J. of Sichuan University Medical Science Edition* **04**

[4] Zhang HJ, Yin F, Chen ML, Qi AQ, Yang LY, Cui WW, Yang SS, Wen G 2021 Six-Class Classification of Leukocytes Based on Machine Learning. Software *J. of Molecular Imag.* **03**

[5] Sun K, Yao XF, Huang G 2020 Good Salient Object Detection in RGBD Images Based on Sample Selection *Software* **10**

[6] Yao Y, Fu LJ, Ge HJ 2019 Research on Assembly Quality Image Recognition Based on Machine Learning. *Ship Sci. and Tech* **10**

[7] Liu ZY, Liu JL, Zhao P 2020 Machine Learning-Based Image Recognition System for Flowing Caenorhabditis elegans on a Microfluidic Chip. *J. of Electronics Information Technology* **09**

[8] Dong JF 2020 Anomaly Target Detection in Multispectral Ship Images Based on Machine Learning *Beijing Uni. of Chinese Medic.*

[9] Wei XY, Wang JH 2020 Application of Machine Learning in Image Classification of Underground Drainage Pipeline Defects *J. of Jiamusi Uni. Natural Science Edition* **01**

[10] Zhen Z 2021 Application of Machine Learning in Ultrasound Image Recognition of Thyroid Malignant Nodules and Traditional Chinese Medicine Tongue Image Classification [D]. Beijing University of Chinese Medicine *Nanjing Uni. of Infor. Sci. Technology*