

تقييم طرق جمع البيانات Data Collection المهمة 5

إعداد: حسن عبد الرحمن الزهراني
إشراف: د. سائد عدنان.

• البحث محل التقييم: تعزيز تصنيف النفايات البلاستيكية باستخدام الذكاء الاصطناعي.
• الربط بالمهمة السابقة: كطالب باحث، استعرضت في المهمة الرابعة "التصميم المنهجي المختلط" الذي اختاره الباحث. اليوم في هذه المهمة، سأتعلم في الشق العملي: كيف تم فعلياً جمع البيانات (الصور والمقابلات)؟ وكيف تم التأكد من جودتها لتغذية خوارزميات الذكاء الاصطناعي بنجاح؟





Abstract

This research aims to develop an intelligent classification system for plastic waste using artificial intelligence (AI), specifically Convolutional Neural Networks (CNNs), to enhance sorting accuracy and improve recycling efficiency in Saudi Arabia, thereby aligning with Vision 2030 sustainability goals. The study focuses on the unique waste management challenges prevalent in Saudi Arabia, particularly in urban centers like Jeddah, with the goal of creating a system that has practical applications in local recycling facilities. The methodology employed a mixed-methods approach, which included a systematic literature review to identify key research gaps, qualitative interviews with 20 relevant stakeholders to understand local needs, and the subsequent development of an AI-based classification system. This system was trained and validated using a combination of public Kaggle datasets and preliminary local data from the Saudi Investment Recycling Company (SIRC). The results of our experimental models indicate that the proposed system is expected to achieve a classification accuracy of over 95% for plastic waste, a significant improvement over traditional method. Furthermore, the system is projected to reduce operational costs by 25% through optimized collection routes powered by IoT technology and increase recycling efficiency by 30%, which is anticipated to contribute to a 20% reduction in greenhouse gas emissions. To ensure practical implementation and long-term success, the study recommends expanding local data collection efforts, conducting rigorous real-world testing in Saudi recycling facilities, and launching public awareness

وضوح الهدف من جمع البيانات وتناسبه مع أسئلة البحث
الهدف من جمع البيانات :الحصول على بيانات مرئية (صور) لتدريب نماذج التعلم العميق(CNNs) ، وبيانات نوعية لفهم التحديات المحلية في السعودية.
•التناسب مع أسئلة البحث :الهدف متطابق تماماً مع تساؤلات البحث حول كيفية رفع كفاءة التصنيف.

•الشاهد من الرسالة ("This system was trained and validated using a combination of public Kaggle datasets and preliminary local data from the Saudi Investment Recycling Company (SIRC).")

•رأيي كطالب :الهدف واضح جداً، فالاعتماد على بيانات (SIRC) جعل البحث يلامس الواقع السعودي ولا يكتفي بالبيانات المعلبة.



Abstract

This research aims to develop an intelligent classification system for plastic waste using artificial intelligence (AI), specifically Convolutional Neural Networks (CNNs), to enhance sorting accuracy and improve recycling efficiency in Saudi Arabia, thereby aligning with Vision 2030 sustainability goals. The study focuses on the unique waste management challenges prevalent in Saudi Arabia, particularly in urban centers like Jeddah, with the goal of creating a system that has practical applications in local recycling facilities. The methodology employed a mixed-methods approach, which included a systematic literature review to identify key research gaps, qualitative interviews with 20 relevant stakeholders to understand local needs, and the subsequent development of an AI-based classification system. This system was trained and validated using a combination of public Kaggle datasets and preliminary local data from the Saudi Investment Recycling Company (SIRC). The results of our experimental models indicate that the proposed system is expected to achieve a classification accuracy of over 95% for plastic waste, a significant improvement over traditional method. Furthermore, the system is projected to reduce operational costs by 25% through optimized

شمولية الطرق المستخدمة وتفاصيل كيفية تنفيذها
• تنوع الطرق :دمج الباحث بين المقابلات النوعية وجمع الصور لمعالجتها.
• تفاصيل التنفيذ)
المعالجة المسبقة **Preprocessing** - لم يكتفِ الباحث بجمع الصور، بل نفذ خط أنابيب (Pipeline) دقيق لمعالجتها.

• الشاهد من الرسالة (المقابلات) ("qualitative : interviews with 20 relevant stakeholders to understand local needs").
• الشاهد من الرسالة (تنفيذ معالجة الصور) ("All : images were resized to a fixed resolution of 224x224 pixels... converted from the BGR color space... to RGB").

دقة تحديد العينة وطريقة اختيارها

• وصف العينة (حجم البيانات المرئية): حدد الباحث حجم عينة الصور بدقة بالغة وقسمها علمياً إلى بيانات تدريب وبيانات اختبار.

• الشاهد من الرسالة 10,617 and 22,564 images for training and validation, classified into two primary categories: Organic (O) and Recyclable (R)".

• التعليق كمتعلم: هذا التحديد الدقيق يعكس فهماً عميقاً لمتطلبات خوارزميات الذكاء الاصطناعي التي تحتاج لعينة ضخمة ومقسمة بعناية لتجنب مشكلة حفظ البيانات (Overfitting).

4.2 Data Preparation Methodology

The quality and preparation of data are foundational to the success of any deep learning model. A meticulous methodology was followed to ensure that the data fed into the models was uniform, high-quality, and capable of enhancing the models' generalization capabilities.

4.2.1 Dataset Description and Analysis

The study utilized a publicly available image dataset specifically curated for waste classification, pre-partitioned into training and testing sets. The dataset is designed for a binary classification task, distinguishing between two principal categories:

- Organic (O): Encompassing waste of biological origin, such as food scraps.
- Recyclable (R): Including materials that can be reprocessed, such as plastics and paper.

A slight class imbalance was observed within the training set, with organic waste constituting approximately 58% of the images. While not severe, this imbalance was effectively addressed through data augmentation techniques to prevent the model from developing a bias towards the majority class.

1. Image Dimensionality Unification: All images were resized to a fixed resolution of 224x224 pixels. This is a mandatory procedure for batch processing and conforms to the fixed input size expected by the pre-trained architectures like ResNet.
2. Color Space Conversion: Images were converted from the BGR color space (default in OpenCV) to RGB, the standard format on which the PyTorch models were pre-trained.
3. Standardization: Pixel values were first normalized to the [0, 1] float range and subsequently standardized using the ImageNet mean [0.485, 0.456, 0.406] and standard deviation [0.229, 0.224, 0.225]. This step is obligatory for transfer learning to ensure that the statistical distribution of the input data is similar to that of the original data the model was trained on.
4. Data Augmentation: To prevent overfitting and enhance the model's ability to generalize to new, unseen images, a series of random transformations were applied to the training data only:
 - RandomHorizontalFlip & RandomVerticalFlip: These transformations introduce orientation invariance, teaching the model to recognize objects regardless of their orientation.
 - RandomRotation(20): This simulates slight rotational variations, forcing the model to learn the object's intrinsic features rather than its specific pose.

وضوح الجدول الزمني ومراقبة جودة البيانات

• مراقبة الجودة : لاحظ الباحث وجود عدم توازن بسيط في العينة (58% للنفايات العضوية)، وتدخل فوراً لضبط الجودة.

• الشاهد من الرسالة (ضبط الجودة) ("A slight class

imbalance was observed... This imbalance was effectively addressed through data augmentation techniques").

• الشاهد من الرسالة (التوحيد القياسي) ("Standardization:

Pixel values were first normalized to the [0, 1] float range and subsequently standardized using the ImageNet mean").

• الخلاصة : رغم عدم وجود جدول زمني تقليدي (Gantt Chart) ، إلا أن "التسلسل الزمني الإجرائي" لجمع البيانات، ثم معالجتها، ثم توحيدها قياسياً للحفاظ على جودتها، كان درساً عملياً ممتازاً لي في كيفية إدارة بيانات الآلة.



مجموعة جامعة
الجميع الذكية
للتعليم القابضة

MIDOCEAN
UNIVERSITY



Thank You

