Orange birds with wings and a heart with text

Description automatically generated with medium confidence

School of Consumer Intelligence and

Information Systems

Department of Applied Information Systems

4. Data Collection & Analysis

Research title: Harnessing Deep Learning For Predicting Traffic Congestion Using Image Classification

KENAN MALALE KASONGO – 220024976

SUPERVISOR: dr tebogo bokaba

Table of Contents

[4. Data Collection 3](#_Toc177243962)

[4.1 Introduction 3](#_Toc177243963)

[4.2 Dataset Selection 3](#_Toc177243964)

[4.3 Data Security and Accessibility 4](#_Toc177243965)

[4.4 Ethical Considerations 4](#_Toc177243966)

[4.5 Conclusion 5](#_Toc177243967)

# Data Collection

# Introduction

Data collection is a critical component of this research, forming the empirical basis for developing a deep learning model to predict road traffic congestion through image classification (Abdullah et al., 2023). This section outlines the effective and thorough implementation of data collection methods.

# Dataset Selection

The dataset selected for this study is the **Traffic-Net** dataset, publicly available on Kaggle. This dataset was chosen due to its direct relevance to the research objectives, images, and balanced representation of various traffic conditions and incidents that impact road congestion.

**Dataset Composition**:

* **Total Images**: 4,400 images.
* **Categories**:
  + Accident
  + Dense Traffic
  + Fire
  + Sparse Traffic
* **Training Set**: 3,600 images, with 900 images in each category.
* **Test Set**: 800 images, with 200 images in each category.

The images encompass a variety of environmental conditions, including different weather scenarios and lighting conditions, enhancing the model's ability to generalize to real-world situations.

**Preprocessing Steps**:

* All images will be resized to **224x224 pixels**.
* Images will be **normalized** to a pixel value range of **[0,1]**.
* **Data augmentation**: random flips, rotations, and lighting adjustments will be applied to improve the model's generalization ability and performance on unseen data.

# Data Security and Accessibility

For added transparency, the data and any supporting resources, such as code or supplementary information, will be kept in a GitHub long-term repository. Access can be granted to any individual who requests it, and a link to the repository will be included in the final research documentation. This approach guarantees that the research may be expanded upon in the future and contributes to the body of scientific knowledge.

By utilizing GitHub, version control is maintained, allowing for tracking changes and collaborative opportunities. Data security is ensured through GitHub's secure infrastructure, and access permissions can be managed to control who can view or contribute to the repository. This strategy effectively balances the need for data security with the accessibility required for academic research.

# Ethical Considerations

Ethical compliance was maintained throughout the data collection process:

* **Privacy Protection:**
  + **Anonymity Assurance:** The images do not contain personally identifiable information, such as recognizable faces or license plates, thereby safeguarding individual privacy.
  + **Compliance with Regulations:** Adhered to the General Data Protection Regulation (GDPR) and institutional ethical standards.
* **Responsible Use:**
  + **Licensing Compliance:** The dataset is used strictly within the bounds of the Kaggle License Agreement, exclusively for educational and research purposes.
  + **Attribution:** Proper credit is given to the dataset creator in all reports and publications resulting from this research, adhering to academic standards and intellectual property rights.

# Conclusion

The data collection process was executed with meticulous attention to detail, ensuring effective implementation of methods that align with the research methodology. By rigorously following established protocols and minimizing errors, high data quality was maintained. The data was effectively secured and organized, ensuring both security and accessibility through the use of a GitHub repository. This robust data foundation is essential for the subsequent development and evaluation of the deep learning model for predicting road traffic congestion.

# Data Analysis

Abdullah, S. M., Periyasamy, M., Kamaludeen, N. A., Towfek, S. K., Marappan, R., Kidambi Raju, S., Alharbi, A. H., & Khafaga, D. S. (2023). Optimizing Traffic Flow in Smart Cities: Soft GRU-Based Recurrent Neural Networks for Enhanced Congestion Prediction Using Deep Learning. *Sustainability*.