UNIVERSIDADE ESTADUAL DE MARINGÁ
CENTRO DE TECNOLOGIA
DEPARTAMENTO DE ENGENHARIA DE PRODUÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA DE PRODUÇÃO

Masters Dissertation

INÊS BARÃO FERREIRA MIYAMOTO

Explainable Deep Learning Applied to Plastics Sorting based on Computer Vision

Maringá
2023

INÊS BARÃO FERREIRA MIYAMOTO

Explainable Deep Learning Applied to Plastics Sorting based on Computer Vision

Dissertation presented to the Postgraduate Program in Production Engineering of the Production Engineering Department, Technology Center of the State University of Maringá (Programa de Pós-Graduação em Engenharia de Produção do Departamento de Engenharia de Produção, Centro de Tecnologia da Universidade Estadual de Maringá), as a partial requirement for the title of Master in Production Engineering.
Area of Concentration: Production Engineering

Supervisor: Prof. Dr. Rodrigo Clemente Thom de Souza
Co-supervisor: Prof.ª Dr.ª Gislaine Camila Lapasini Leal

Maringá
2023

Dados Internacionais de Catalogação-na-Publicação (CIP)
(Biblioteca Central - UEM, Maringá - PR, Brasil)

Miyamoto, Inês Barão Ferreira

M685e Explainable deep learning applied to plastics sorting based on computer vision / Inês Barão Ferreira Miyamoto. -- Maringá, PR, 2023.
78 f.: il. color., figs., tabs.

Orientador: Prof. Dr. Rodrigo Clemente Thom de Souza.
Coorientadora: Profa. Dra. Gislaine Camila Lapasini Leal Leal.
Dissertação (Mestrado) - Universidade Estadual de Maringá, Centro de Tecnologia, Departamento de Engenharia de Produção, Programa de Pós-Graduação em Engenharia de Produção, 2023.

1. Visão computacional. 2. Inteligência artificial. 3. Triagem de plásticos - Reciclagem. I. Souza, Rodrigo Clemente Thom de, orient. II. Leal, Gislaine Camila Lapasini Leal, coorient. III. Universidade Estadual de Maringá. Centro de Tecnologia. Departamento de Engenharia de Produção. Programa de Pós-Graduação em Engenharia de Produção. IV. Título.

CDD 23.ed. 006.3

Márcia Regina Paiva - CRB-9/1267

FOLHA DE APROVAÇÃO

INÊS BARÃO FERREIRA MIYAMOTO

Explainable Deep Learning Applied to Plastics Sorting based on Computer Vision

Dissertação apresentada ao Programa de Pós-Graduação em Engenharia de Produção do Departamento de Engenharia de Produção, Centro de Tecnologia da Universidade Estadual de Maringá, como requisito parcial para obtenção do título de Mestre em Engenharia de Produção pela Banca Examinadora composta pelos membros:

BANCA EXAMINADORA

Prof. Dr. Rodrigo Clemente Thom de Souza
Presidente Orientador
Universidade Estadual de Maringá – PGP/UEM

Prof. Dr. Rafael Henrique Palma Lima
Membro examinador interno
Universidade Estadual de Maringá – PGP/UEM

Profa. Dra. Juliana Verga Shirabayashi
Membro examinadora externo
Universidade Federal do Paraná – UFPR

Aprovada em: 08 de fevereiro de 2023.
Local da defesa: Sala de Projeção, Bloco 19, *campus* da Universidade Estadual de Maringá.

ACKNOWLEDGMENTS

First, I thank God, who guided me and gave me strength during this arduous journey.
I would also like to thank my husband, for the unconditional support from the beginning, and for helping me whenever he could.
A special thanks to my supervisor, for always guiding me and helping me overcome difficulties whenever I needed, and for always being so understanding.
I also thank my co-supervisor for helping me when necessary and for the contributions.
Also, a big thank you to the members of the examination board, professors Rafael and Juliana, who provided excellent contributions to my dissertation during the qualification exam.
I also sincerely thank João and Jorge, who greatly contributed to the elaboration of the articles that compose this dissertation.
I thank CAPES, for the financial support.
And to all the people that somehow contributed and support me during this journey, I am deeply grateful.

EPIGRAPH

The time will come when we shall realize that all we have paid has been nothing at all by comparison with the greatness of our prizes.

(SAINT TERESA OF AVILA)

Explainable Deep Learning Applied to Plastics Sorting based on Computer Vision

ABSTRACT

Recycling plastics is essential to reduce the environmental impact and resource depletion, as it can decrease energy and material usage per unit of output and so yield improved eco-efficiency. Sorting plastics is one of the stages of the recycling process, and it is necessary because there are different types of plastics, and recycling processes generally require a single polymer. There are various techniques for sorting plastic materials, including manual techniques, which are time-consuming and not very efficient, and Machine Learning (ML) and Deep Learning (DL) techniques, which have been producing good results, but they are black boxes. That means those models are not transparent and, consequently, they are not reliable. Explainable Artificial Intelligence (XAI) is a recent research topic that has been producing interesting results in the last few years and it is considered a solution to overcome those limitations in classical DL methods. Thus, this work aims to combine and apply DL and XAI techniques to plastics sorting. First, a systematic literature review (SLR) was conducted on XAI applied to DL, to identify the main DL and XAI techniques in the literature, as well as the metrics to assess XAI techniques, the limitations and future research suggestions. Afterwards, CNNs, ResNet50, ResNet152, VGG19, and VGG16, and XAI techniques, Integrated Gradients and Guided Grad-CAM, were applied to the data images. Finally, DL metrics, accuracy, precision, recall, and F1-score, and XAI metrics, Max-Sensitivity and Infidelity, were employed to compare the methods. The results reveal that ResNet152 achieved the best performance, and Guided Grad-CAM produced better explanations, in general, although not for ResNet152. This dissertation is expected to further advance research on XAI applied to DL and contribute to the recycling of plastics, specifically to the essential step of sorting them.

***Keywords*:** Deep Learning. Explainable Artificial Intelligence. Computer Vision. Plastics Sorting.

Aprendizagem Profunda Explicável aplicada à Triagem de Plásticos baseada em Visão Computacional

RESUMO

A reciclagem de plásticos é essencial para reduzir o impacto ambiental e o esgotamento de recursos, uma vez que pode diminuir a utilização de energia e material por unidade de produção e, assim, produzir uma maior ecoeficiência. A separação de plásticos é uma das etapas do processo de reciclagem, e é importante pois existem diferentes tipos de plásticos e os processos de reciclagem requerem geralmente um único polímero. Existem várias técnicas de triagem de materiais plásticos, incluindo técnicas manuais, que consomem muito tempo e não são muito eficientes, e técnicas de Machine Learning (ML) e Deep Learning (DL), que têm vindo a produzir bons resultados, mas são caixas pretas. Isto significa que estes modelos não são transparentes e, consequentemente, não são confiáveis. A Inteligência Artificial Explicável (XAI) é um tema recente que tem produzido resultados interessantes nos últimos anos e é considerada uma solução para ultrapassar essas limitações nos métodos clássicos de DL. Assim, este trabalho visa combinar e aplicar técnicas de DL e XAI à triagem de plásticos. Primeiro, foi realizada uma revisão sistemática da literatura (SLR) sobre XAI aplicada ao DL, para identificar as principais técnicas de DL e XAI na literatura, bem como as métricas para avaliar as técnicas de XAI, as limitações e as sugestões para estudos futuros. Posteriormente, as CNNs ResNet50, ResNet152, VGG19, e VGG16, e as técnicas de XAI *Integrated Gradients* e *Guided Grad-CAM*, foram aplicadas às imagens. Finalmente, foram utilizadas as métricas de DL acurácia, precisão, *recall*, e *F1-score*, e as métricas de XAI *Max-Sensitivity* e *Infidelity*, para comparar os métodos. Os resultados revelam que a ResNet152 alcançou o melhor desempenho, e *Guided Grad-CAM* produziu melhores explicações, em geral, embora não para a ResNet152. Espera-se que esta dissertação contribua para o avanço nas pesquisas sobre XAI aplicada a DL e contribua para a reciclagem de plásticos, especificamente para a etapa essencial da triagem.

Palavras-chave: Aprendizagem Profunda. Inteligência Artificial Explicável. Visão Computacional. Triagem de Plásticos.

LIST OF ILLUSTRATIONS

LIST OF ABBREVIATIONS AND ACRONYMS

AF	*Atrial Fibrillation*
AI	*Artificial Intelligence*
ANN	*Artificial Neural Network*
AVE	*Average Euclidean Distance*
CAE	*Convolutional Autoenconder*
CASTLE	*Cluster-aided space transformation for local explanations*
ChIMP	*Choquet Integral Multilayer Perceptron*
CIE	*Confident Itemsets Explanation*
CK	*Cohen-Kappa*
CLRP	*Contrastive Layer-wise Relevance Propagation*
CMIE	*Customizable Model Interpretation Evaluation*
CNC	*Computer Numerical Control*
CNN	*Convolutional Neural Network*
CX-ToM	*Counterfactual Explanations – Theory of Mind*
DL	*Deep Learning*
DNN	*Deep Neural Networks*
DRL	*Deep Reinforcement Learning*
EBPG	*Energy-Based Pointing Game*
EG	*Effective Gradient*
ELM	*Extreme Learning Machines*
ETeMoX	*Event-driven Temporal Models for Explanations*
FC	*Fully Connected*
FIA	*Faithfulness, interpretability and applicability*
FOX	*Neuro-Fuzzy model for process Outcome prediction and eXplanation*
GA	*Geographic Atrophy*
GBP	*Guided BackPropagation*
HDPE	*High-Density Polyethylene*
IG	*Integrated Gradients*
I*G	*Input * Gradient*
IoU	*Intersection-over-Union*
IRTEX	*Image Retrieval with Textual Explanations*
LDPE	*Low-Density Polyethylene*
LGNN	*Locality Guided Neural Network*
LIME	*Local Interpretable Model-Agnostic Explanations*

LRP	*Layer-wise Relevance Propagation*
LSTM	*Long Short-Term Memory*
ML	*Machine Learning*
MoRF	*Most Relevant First*
MS	*Maximum Sensitivity/Max-Sensitivity*
PA	*Pixel Accuracy*
PACE	*Post-hoc Architecture Agnostic Concept Extractor*
PET	*Polyethylene Terephthalate*
PP	*Polypropylene*
PPM	*Predictive Process Monitoring*
PS	*Polystyrene*
PVC	*Polyvinylchloride*
ResNet	*Residual Neural Network*
RAP	*Relative Attributing Propagation*
RBM	*Restricted Boltzmann Machines*
ReLU	*Rectified Linear Units*
RISE	*Randomized Input Sampling for Explanation*
RNN	*Recurrent Neural Network*
ROE	*Region of Evidence*
RQ	*Research Questions*
RT	*Response Time*
SAE	*Stacked Autoencoders*
SGLRP	*Softmax Gradient Layer-wise Relevance Propagation*
SHAP	*Shapley Additive Explanations*
SLR	*Systematic Literature Review*
SSim	*Structural Similarity*
XAI	*Explainable Artificial Intelligence*
XAUG	*eXpert AUGmented*

TABLE OF CONTENTS

1

INTRODUCTION

Sustainable development is increasingly presented as necessary to all that is good and desirable in society. The term sustainability has its origin in ecological science and was developed to express the conditions that must exist for the ecosystem to sustain itself over the long term (HOLDEN; LINNERUD, 2007). In the Brundtland Report (WCED, 1987), there are several references to the necessity of ecological sustainability.

A few centuries of industrial economic development, with limited attention for environmental and social externalities, have resulted in challenges to the environment, such as excessive air and water pollution, deforestation, climate change, overpopulation, migration, massive poverty, social inequality, and natural resource scarcity (HUMMELS; ARGYROU, 2021).

People generate solid waste through their daily activities, which needs to be properly managed in a way that minimizes risk to the environment and human health (UITERKAMP; AZADI; HO, 2011). Empirical evidence shows that recycling waste is environmentally and economically efficient, as it reduces damage to the environment and also saves energy, preserves resources and saves waste collection and disposal costs (KASEVA; GUPTA, 1996).

Plastic is a versatile and strong material, and its production has increased nearly tenfold since 1950, for a variety of applications (GEYER *et al*., 2017). Recycling specifically plastics has been increasingly necessary, both for economic and environmental reasons (SCOTT, 1995). One of the reasons is that, in the consumer societies such as Europe and America, scarce petroleum resources are used for the production of various types of plastics for an even wider

variety of products (NKWACHUKWU *et al*., 2013). Development of synthetic polymers, used to make plastics such as polyethylene, polypropylenes, polyesters and polyamides, has revolutionized the types of containers for products, the types of materials for packaging and other products made of plastic (NKWACHUKWU *et al*., 2013). However, most of these polymers are not biodegradable; therefore, when they are used and discarded, they can become waste and pollute the environment for a very long time, which may be harmful to human health and the environment (SARDON; DOVE, 2018). Therefore, technology to economically recover these polymeric materials and return them into the materials supply chain is necessary (AYRE, 2018).

Recycling plastics is a way to reduce environmental impact and resource depletion, as it can decrease energy and material usage per unit of output and so yield improved eco-efficiency (HOPEWELL; DVORAK; KOSIOR, 2009). There are various ways of recycling plastics and the ease of recycling depends on the type of polymer, product and package design. For example, rigid containers that consist of a single polymer are simpler and more economic to recycle than multi-layer and multi-component packages (HOPEWELL; DVORAK; KOSIOR, 2009).

Sorting plastics is one of the stages of the recycling process. This separation process is necessary because the presence of even a small quantity of a different type of plastic may decrease the quality of the whole batch (DODBIBA; FUJITA, 2004). So, most sources of recyclable material provide a random mixture of various plastic types, but recycling processes generally require a single polymer to be used (SCOTT, 1995). Some techniques for sorting plastic materials are as follows: wet separating techniques, such as flotation of plastics (SHIBATA *et al*., 1996); dry techniques, such as near-infrared spectroscopic analysis or x-rays (WILLIAMS; NORRIS, 1987); and sorting by melting, which can only be used to separate two plastic types at a time (RUJ *et al*., 2015). There is also the possibility of manually separating the plastics; however, it is not as efficient as automated sorting and tends to take longer. As Jimoh, Ajayi and Ayilara (2014) state, automated sorting systems are necessary in order to achieve high throughput and accuracy.

Besides the aforementioned techniques for separating plastics, there are artificial intelligence and computer vision algorithms for that task. For example, Jimoh, Ajayi and Ayilara (2014) used a fuzzy model to classify images of plastic materials into their respective categories. Additionally, Meeradevi, Raju and Vigneshkumaran (2020) classified plastic bottle images with Convolutional Neural Networks (CNNs), which constitute a type of Deep Learning (DL) model.

Deep Learning (DL) algorithms have been used in a great variety of fields, such as healthcare, manufacturing, autonomous robots and vehicles, cyber-security, sustainability, as well as with many types of data, for example, image processing, classification and detection, speech and audio processing, among others (SARKER, 2021). They can also provide solutions for sorting plastics efficiently, in an automated way (MEERADEVI; RAJU; VIGNESHKUMARAN, 2020; JIMOH; AJAYI; AVILARA, 2014).

DL is a subset of Machine Learning (ML) which typically consists of Artificial Neural Networks (ANN) with more than one hidden layer, organized in deeply nested network architectures (JANIESCH; ZSCHECH; HEINRICH, 2021). Its methods are representation-learning methods with multiple levels of representation, obtained by composing simple but non-linear modules that each transform the representation at one level (starting with the raw input) into a representation at a higher, slightly more abstract level (LECUN; BENGIO; HINTON, 2015).

Although DL models usually have better performance and are more powerful than white-box ones (such as linear models), they are considered black boxes; that means their decisions are hard to understand, with a complex underlying mechanism (LINARDATOS; PAPASTEFANOPOULOS; KOTSIANTIS, 2020). It is difficult to trust systems whose decisions cannot be well-interpreted, especially in sectors such as healthcare or autonomous cars, where moral and fairness issues have naturally arisen (LINARDATOS; PAPASTEFANOPOULOS; KOTSIANTIS, 2020).

Thus, in ML and DL, better approaches have been necessary to effectively comprehend the black box models' decisions, and which characteristics they consider to reach them. Explainable artificial intelligence (XAI) is a recent research topic that has been intriguing in the last few years and it is considered a solution to overcome constraints in classical DL methods. This topic has been producing interesting results, which are possible to observe in recent studies (ADADI; BERRADA, 2018; MURDOCH *et al*., 2019). Some examples of commonly used XAI techniques in the literature are: Shapley Additive Explanations (SHAP), Local Interpretable Model-Agnostic Explanations (LIME), Layer-wise Relevance Propagation (LRP), among others.

As in DL, XAI models can also be assessed by means of different metrics, which are either qualitative (subjective) or quantitative (objective). Qualitative metrics work towards making explanations more human-friendly and meaningful (HOOKER *et al*., 2018). That kind of metric typically verifies if the explanation is aligned with the subject's expectation and how helpful it is in enabling the person to understand the behavior of the model (GILPIN *et al*.,

2018). An example of a qualitative metric is an analysis of the explanations by specialists in the field (VENUGOPAL *et al*., 2020). On the other hand, quantitative metrics assess whether the explanation is a reliable reflection of the model behavior. For instance, Maximum Sensitivity (YEH *et al*., 2019), Infidelity (YEH *et al*., 2019), Most Relevant First (MoRF) (KAKOGEORGIOU; KARANTZALOS, 2021), and the pixel-flipping test (BACH *et al*., 2015) are quantitative metrics. Both approaches are important for evaluating explanations; as Hooker *et al*. (2018) explains: "interpretability methods should be both meaningful to a human and correctly explain model behavior".

In this context, this work aims to combine and apply DL and XAI techniques to plastics sorting. In order to do that, the first step was to conduct a systematic literature review (SLR) on XAI applied to DL, to identify the main DL and XAI techniques in the literature, as well as the metrics to assess XAI techniques, the limitations and future research suggestions; afterwards, based on the SLR results, DL techniques were chosen and employed to classify a dataset from Kaggle, which contained plastic bottle images, into their respective types; and XAI techniques were also defined and applied in order to better understand the classification decisions. Finally, metrics for performance assessment of the XAI techniques were used. Therefore, this work is a multi-paper dissertation composed of two articles: one refers to the SLR, and the other to the application of XAI and DL techniques along with the metrics. Table 1 shows each article that composes this dissertation, including their objectives, methods, and contributions.

Table 1 – Articles contained in this dissertation.

	Article 1	Article 2
Title	Explainable Artificial Intelligence Applied to Deep Learning: A Systematic Literature Review	Explainable Artificial Intelligence Applied to Deep Learning for Plastics Sorting
Objectives	Identify the most used DL and XAI techniques and XAI metrics in the literature	Apply DL and XAI techniques to classify plastic images and use metrics to compare them
Methodology	Databases ScienceDirect, Springer, and IEEExplore, from 2020 and 2021. 108 studies considered relevant and analyzed as to DL and XAI	CNNs (ResNet50, ResNet152, VGG19, VGG16) and XAI (Guided Grad-CAM, Integrated Gradients), assessed by Max-Sensitivity and Infidelity and

	techniques used, area of application, XAI metrics employed, limitations and future research suggestions	applied to public dataset “Plastic Recycling Codes” from Kaggle
Theoretical/Practical Contributions	Advance research on XAI applied to DL and summarize findings of studies on this field from the last few years	Propose an automated way to classify plastics for sorting them during the recycling process

Source: (THE AUTHOR, 2023).

Plastic usually contains a symbol that identifies the resin category, which makes their images suitable to be classified by DL algorithms. However, some of them might be crushed or deformed and that might result in a misclassification; that is why XAI techniques are helpful, as they make the model more transparent and, consequently, more reliable. Furthermore, in literature, there is a scarcity of studies that applied ML or DL techniques to plastics recycling and sorting, whereas studies combining DL and XAI in that type of problem were not found at all. That is the novelty this work presents, since it aims to fill that gap and contribute to the literature in that regard.

The next sections of this work are as follows: Section 2 contains the first article, the SLR; Section 3 is the second article; and Section 4 presents a general conclusion.

2

ARTICLE 1: EXPLAINABLE ARTIFICIAL INTELLIGENCE APPLIED TO DEEP LEARNING: A SYSTEMATIC LITERATURE REVIEW

This section refers to the first article, entitled Explainable Artificial Intelligence Applied to Deep Learning: A Systematic Literature Review.

2.1 INTRODUCTION

Deep Learning (DL) is a subset of Machine Learning (ML) which typically consists of Artificial Neural Networks (ANN) with more than one hidden layer, organized in deeply nested network architectures (JANIESCH; ZSCHECH; HEINRICH, 2021). DL allows computational models that are composed of multiple processing layers to learn representations of data with multiple levels of abstraction (LECUN; BENGIO; HINTON, 2015). Its methods are representation-learning with multiple levels of representation, obtained by composing simple but non-linear

modules that each transform the representation at one level (starting with the raw input) into a representation at a higher, slightly more abstract level (LECUN; BENGIO; HINTON, 2015).

DL algorithms have been used in healthcare, ophthalmology, autonomous robots and vehicles, image processing, classification and detection, speech and audio processing, cyber-security and many other areas and applications. This indicates the reach of DL algorithms in our daily lives (DAS; RAD, 2020).

Although DL models usually have better performance and are more powerful than white-box ones (such as linear models), they are considered black boxes; therefore, they are hard to understand, with a complex underlying mechanism (LINARDATOS; PAPASTEFANOPOULOS; KOTSIANTIS, 2020).

Systems whose decisions cannot be well-interpreted are difficult to be trusted, especially in sectors, such as healthcare or self-driving cars, where moral and fairness issues have naturally arisen. As Ras et al. (2022) stated, trust and justification can hardly be achieved if the user is not provided with an adequate explanation for the process that generated the recommendation. For example, an error in a medical system is a danger to human lives, and that is the reason why doctors have expressed concern about the black-box nature of DL algorithms (JIA; REN; CAI, 2020).

Thus, there has been a need for better approaches to effectively comprehend the black box models' decisions. Explainable artificial intelligence (XAI) is a recent research topic that has been arising in the last few years and it is thought to be a solution to overcome constraints in classical DL techniques. This topic has been producing interesting results, shown in recent studies (ADADI; BERRADA, 2018; MURDOCH et al., 2019).

XAI encompasses Machine Learning (ML) or AI systems/tools for demystifying black models, that is, what the models have learned, and/or for explaining individual predictions (SAMEK; WIEGAND; MÜLLER, 2017). Explanations should improve human understanding and confidence in decision making, as this is necessary to make the models more reliable (SAMEK; WIEGAND; MÜLLER, 2017).

Explainable DL is a relatively recent topic, as we can see in Figure 1 below; knowing that, we consider it important to summarize the most used techniques and performance metrics in the literature, as well as to understand the main limitations and suggestions for the future. Therefore, this study is a systematic literature review on the application of XAI in DL, with the aim of analyzing the contribution of XAI techniques in explaining decisions of DL algorithms, over the last two years.

The sections of this article are organized as follows: this introduction constitutes the first section; section 2 refers to the methodology of this review, and encompasses the research questions, inclusion and exclusion criteria, as well as research sources and strings; section 3 presents the results, with each subsection (3.1. to 3.5.) presenting the answers to the research questions; and, finally, section 4 is the conclusion, which contains limitations and suggestions for future research.

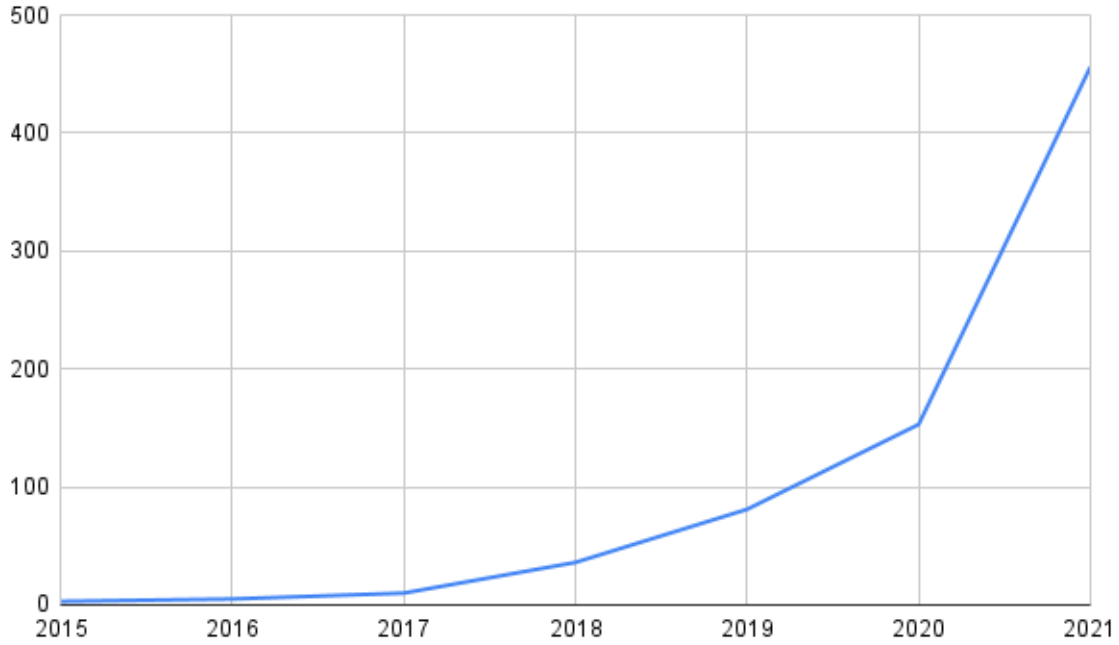
2.2 METHODOLOGY

In this section, the methodology of this systematic literature review is presented, which includes the stages of planning (section 2.2.1) and conducting (section 2.2.2). Planning refers to the process of preparing the review, such as defining the strings and databases (SIDDAWAY; WOOD; HEDGES, 2019). Conducting refers to the research in the databases, the articles found and those considered relevant (SIDDAWAY; WOOD; HEDGES, 2019). Afterwards, the relevant articles were analyzed.

2.2.1 Planning

The first step was to define the research databases, which were: ScienceDirect, Springer and IEEExplore. These databases are the most relevant for the area and topic of this research. Afterwards, the strings for the research were also selected: “(xai OR explainable artificial intelligence) AND (deep learning OR image classification OR neural networks OR computer vision)”. As to the period of time, we considered studies from 2020 and 2021, as XAI is a very recent research topic and very few studies were published before 2020. In fact, the majority of studies on XAI applied to DL are from 2021. This is shown in Figure 1.

Figure 1 - Number of publications on XAI applied to DL, from 2015 to 2021.



Source: (THE AUTHORS, 2022).

The research questions (RQ's) formulated according to the aim of this study are as follows:

RQ1: "What are the main XAI techniques applied to DL recently?";

RQ2: "What are the DL techniques in the context of XAI that have been used recently?";

RQ3: "What are the recent DL applications in the context of XAI?";

RQ4: "What kind of metrics have been used to assess XAI techniques in DL?"; and

RQ5: "What are the main limitations and future research suggestions?".

The research in the established databases was conducted on January 20, 2022. Afterwards, a list of the articles found was generated. Based on that list, a quick overview of the content in each of the articles was done, so as to determine their relevance to this literature review, and whether they should be used or otherwise discarded. In order to do that, the titles and keywords of the articles were analyzed; if they were considered relevant, then the abstracts went through the same procedure; and finally, the introduction and conclusion. Finally, the articles considered relevant were thoroughly read.

Regarding the inclusion/exclusion criteria, only articles in English were included. Furthermore, articles that did not address both DL and XAI were excluded, as well as those that did not specifically apply XAI to DL (for example, other literature reviews and theoretical studies). Also, books and book chapters were not included; only articles published in journals or conferences. Table 1 contains the inclusion (I) and exclusion (E) criteria considered for this

review.

Table 1 - Inclusion and Exclusion Criteria

Type	Criteria
I	Studies in English
I	Studies published as journal or conference articles
I	Studies that address XAI applied to DL
E	Studies in other languages besides English
E	Literature Reviews and other theoretical studies
E	Books or book chapters
E	Studies that do not address both XAI and DL

Source: (THE AUTHORS, 2022).

2.2.2 Conducting

After conducting the research on the established databases, the number of articles found is shown in Table 2.

Table 2 - Number of articles, total and relevant, on each database.

	Springer	ScienceDirect	IEEExplore	Total
Total	244	270	95	609
Relevant	19	47	42	108

Source: (THE AUTHORS, 2022).

A large number of studies found on the Springer database were not considered relevant, since they only briefly mentioned XAI. The same happened on ScienceDirect, especially with the results that appeared last in the search.

A reason for this might be that searching for the strings in the title, abstract and keywords was not possible. In order not to miss any articles that might be relevant, we searched the entire text and then thoroughly analyzed the studies. Therefore, the word “XAI” appeared many times in the discussion and/or conclusion, as a suggestion for future research, for example, but was not the main topic, or among the topics, of the article.

Other articles, such as Literature Reviews, had a more theoretical nature, so they did not contain a specific contribution of XAI to DL. Besides, some articles were not specifically about DL, but ML. Therefore, they were not relevant to our review, either.

Besides, many conference articles on Springer were published as book chapters. They were excluded from this review, as were all books in the search results.

Appendix A contains a table with the selected articles.

2.3 RQ1: XAI TECHNIQUES

Although the definition of “explanation” varies, in its most general form, an explanation is any information that can help the user understand and communicate to others why the model exhibits a particular pattern of decision-making and how individual decisions come about (RAS et al., 2021). Explanations can play different roles, such as giving insight into model training and generalization or into model predictions. The latter is the most usual and helps practitioners explain why the model made a particular prediction, usually in terms of the model input (RAS et al., 2021).

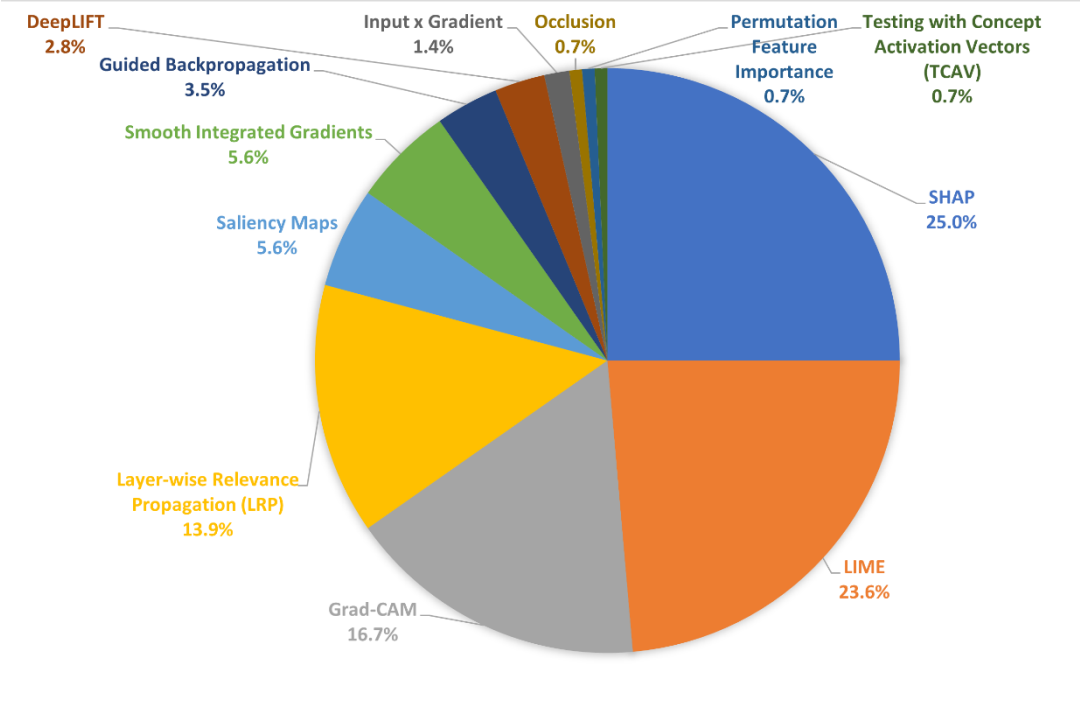
XAI techniques can be categorized according to their scope, method and use. The scope of the explanation, global to the model versus local to the prediction, corresponds to reliability at two levels - trust in the model versus trust in the prediction (RIBEIRO; SINGH; GUESTRIN, 2016). An example of a method for local predictions is CASTLE, a novel technique (LA GATTA et al., 2021).

As to the use, explanations can be model-specific or model-agnostic (post-hoc). Model-specific techniques incorporate interpretability constraints within the inherent structure and learning mechanisms underlying deep learning models, whereas model-agnostic techniques use the inputs and predictions of the black box models to generate explanations (RAI; 2020). Confident Itemsets Explanation (CIE) for post-hoc explanations was applied to an MLP (Multilayer Perceptron) (MORADI; SAMWALD, 2021). SHAP was applied to provide explanations to the decisions made by a DRL agent (LØVER et al., 2021).

Furthermore, the techniques can also be classified according to the type of method: gradient-based or perturbation-based. Gradient-based techniques are designed to explain predictions f(x) for any $x \in X$, by computing the derivative of f(x) with respect to each feature of x (ANCONA et al., 2018). Examples of gradient-based techniques are SmoothGrad, Integrated Gradients and Grad-CAM. LIME is a perturbation-based technique (RIBEIRO; SINGH; GUESTRIN, 2016) and has been used in many studies (NEVES et al., 2021; KINKEAD et al., 2021; SCHÖNHOF et al., 2021).

Figure 2 contains a chart with the percentage of use of the XAI techniques, and is represented below

Figure 2 - Percentage of use of each XAI technique



Source: (THE AUTHORS, 2022).

In the articles considered in this review, the techniques most commonly applied to DL models were SHAP (25%) and LIME (23.6%). These are perturbation-based techniques and usually for local explanations. Grad-CAM and LRP have been frequently applied as well. An example is the study by Jung et al. (2021), who propose selective LRP, which produces a clearer heatmap than the existing techniques by combining relevance-based techniques and gradient-based techniques, and is applied to CNNs and RNNs.

It is important to highlight that many studies have employed more than one XAI technique, especially to compare their performances. Kakogeorgiou and Karantzalos (2021) evaluated the performance of ten different XAI techniques, and found Occlusion, Lime and Grad-CAM techniques were the most interpretable and sensitive.

This finding is in line with the study by Lee et al. (2021a), who state that visualization techniques are the most common XAI techniques, when applied to DL, and they include LIME, CAM, LRP and Guided Backpropagation. Although not stated in this excerpt, SHAP can also be considered a visualization technique. López-Cabrera et al. (2021) reached a similar conclusion and found that, among the most used XAI techniques, are LIME, Grad-CAM and Grad-CAM++.

Regardless of the technique, what stood out the most is that all studies obtained good results from the application of XAI techniques to black-box models, despite finding some limitations, detailed in section 3.5.

2.3.1 Novel XAI Techniques

Due to the novelty of XAI, some authors have found it necessary to introduce new techniques for explainability of black-box models and test them in their studies. Thus, some of the articles analyzed did not use state-of-the-art XAI techniques, such as LIME or SHAP, but instead proposed new ones. Those are not included in the chart that represents the percentage of use of the techniques.

La Gatta et al. (2021) proposed a novel XAI technique, CASTLE, for local explanations. The technique can be used in any area, and it was tested on six different datasets containing data related to various fields. Although it was meant for local explanations, the authors concede that combining local and global explanations is likely to provide better insight into the model.

Moradi and Samwald (2021) proposed an explanation technique named Confident Itemsets Explanation (CIE) for post-hoc explanations, and used it together with MLP. The technique produces instance-wise and class-wise explanations that accurately approximate the behavior of the target black-box.

In an attempt to introduce XAI into Predictive Process Monitoring (PPM), the authors Pasquadibisceglie et al. (2021) proposed a fully interpretable model for outcome prediction - FOX. The proposed technique is based on a set of fuzzy rules acquired from event data via the training of a neuro-fuzzy network. The authors claim this solution provides a good trade-off between accuracy and interpretability of the predictive model.

Some authors simply embedded explainability into the DL models, without giving the technique a name. Jo et al. (2020) conducted a study in which an explainable DLM was developed using ECGs and then internally and externally validated. To do that, they developed modules to classify the characteristics of Atrial Fibrillation (AF), not its presence. The two modules developed for feature and final ensemble DLM used three labels of each ECG based on supervised learning: one to to determine the irregularity of heart rhythm, another to determine the absence of p-wave, and finally, concatenated the two modules of features and developed a final explainable DLM to detect AF. The results indicated that the XAI methodology could be used to describe the reason for the decision made by the DLM in detecting AF with high performance.

2.4 RQ2: DL TECHNIQUES

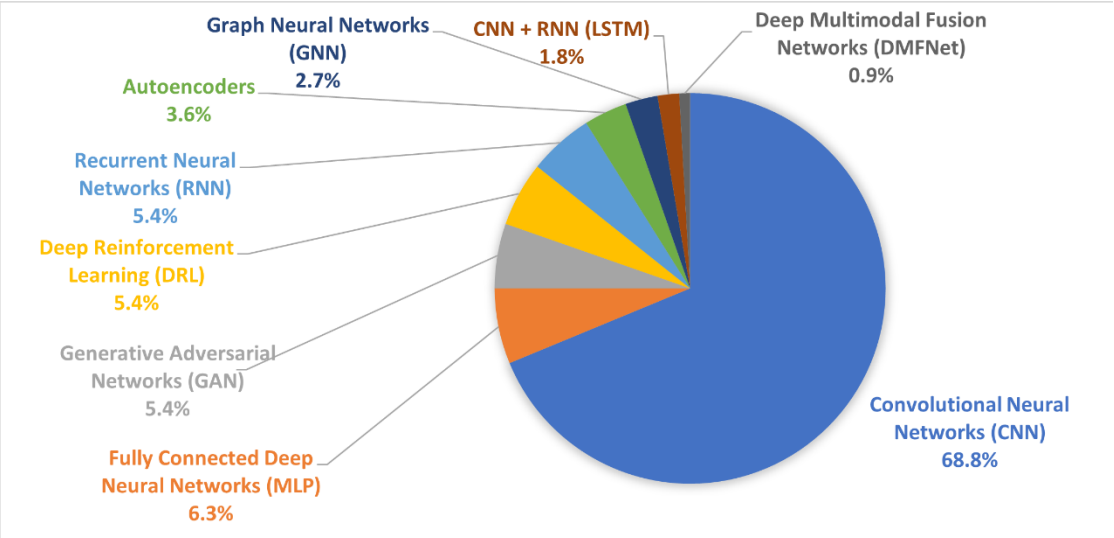
Deep Learning (DL) is a subset of Machine Learning (ML) which typically consists of Artificial Neural Networks (ANN) with more than one hidden layer, organized in deeply nested network architectures (JANIESCH; ZSCHECH; HEINRICH, 2021). It has been used successfully in tasks such as image classification and object recognition, and in many areas, especially in medicine, as seen in the last section.

DL is particularly useful when dealing with large and high-dimensional data, which is why DNNs tend to outperform shallow ML algorithms for most applications in which text, image, video, speech, and audio data needs to be processed (LECUN *et al.*, 2015). However, for low-dimensional data input, especially in cases of limited training data availability, shallow ML can still produce superior results, which even tend to be better interpretable than those generated by deep neural networks (RUDIN, 2019). Consequently, DL can highly benefit from XAI.

There are many DL techniques in the literature. Some of the best known are: DBNs (Deep Belief Networks), RBMs (Restricted Boltzmann Machines), CNNs (Convolutional Neural Networks), RNNs (Recurrent Neural Networks) or LSTMs (Long Short-Term Memory), SAE (Stacked Autoencoders) and DRL (Deep Reinforcement Learning) (SHAMSHIRBAND; RABCZUK; CHAU, 2019).

In order to understand which DL techniques have been used the most, which is the aim of RQ2, we generated a chart containing the percentages of use of each technique. The chart, is shown below, in Figure 3.

Figure 3 - Percentage of use of each Deep Learning technique



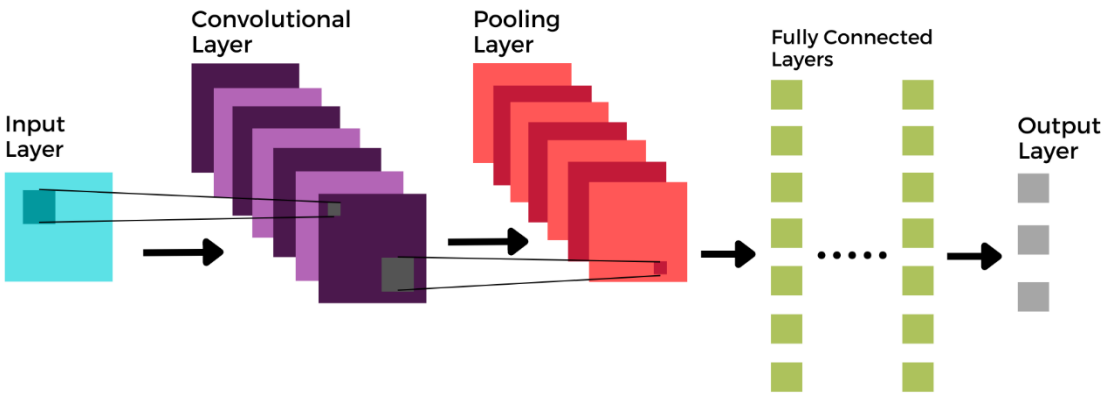
Source: (THE AUTHORS, 2022).

It is possible to observe that CNNs have been the most used DL technique (68.8%), combined with XAI. They are, indeed, referenced among the most used techniques in other literature reviews (SHAMSHIRBAND; RABCZUK; CHAU, 2019; VOULODIMOS *et al*., 2018). The reason for that is likely because, as Sarker (2021) states, "this commonly used technique [CNNs] is a very efficient technique in various applications, especially computer vision". Hassanzadeh, Essam and Sarker (2022) state that CNN has been one of the most relevant image classification techniques in the last few years. Over the years, a diversity of research articles has proven those statements to be generally true.

A CNN is a type of Deep Neural Network (DNN) that was introduced by LeCun *et al*. in 1989 (LECUN *et al*., 1998). In general, a CNN consists of a stack of modules that perform three operations on the input: convolution, rectified Linear Units (ReLU), and pooling. To briefly explain each of them, during the Convolution, filters are applied to the input image (or previous layer) to produce a new layer called the output layer, which may have a different height, width, or depth than the input layer (AGARWAL; GUDI; SAXENA, 2020). The output layer is computed by sliding the filters across the input and performing element-wise multiplication and, afterwards, all the resultant layers are added to produce a single feature map (AGARWAL; GUDI; SAXENA, 2020). Then, The ReLU function is an activation function, the most used in CNNs, which is applied to introduce non-linearity after the convolution operation. After ReLU, in pooling, the image is downsampled by preserving only the maximum or the average of values in a neighborhood - this is done to reduce computational time (AGARWAL; GUDI; SAXENA, 2020). Finally, we have fully connected (FC) layers in which every neuron in the input layer is connected to every neuron in the output layer. The FC layers are used to classify the images based on the features extracted by the convolutional layers (AGARWAL; GUDI; SAXENA, 2020).

Figure 4 below shows the general architecture of a CNN.

Figure 4 - General architecture of a CNN.



Source: (THE AUTHORS, 2022).

Fully-connected DNNs, specifically MLP (MultiLayer Perceptron), are also present in a significant number of articles. A typical MLP is a fully connected network that consists of an input layer, an output layer that makes a decision about the input, and one or more hidden layers between these, considered as the network's computational engine (SARKER, 2021). Behl *et al*. (2021) applied MLP for the multi-class classification of Twitter data related to COVID-19 and other disasters, and compared it with CNNs. They found MLP yielded better results in all the test cases and proved to be usable for data from an unseen disaster. Also, Moradi and Samwald (2021), which proposed the XAI technique CIE, used it together with MLP.

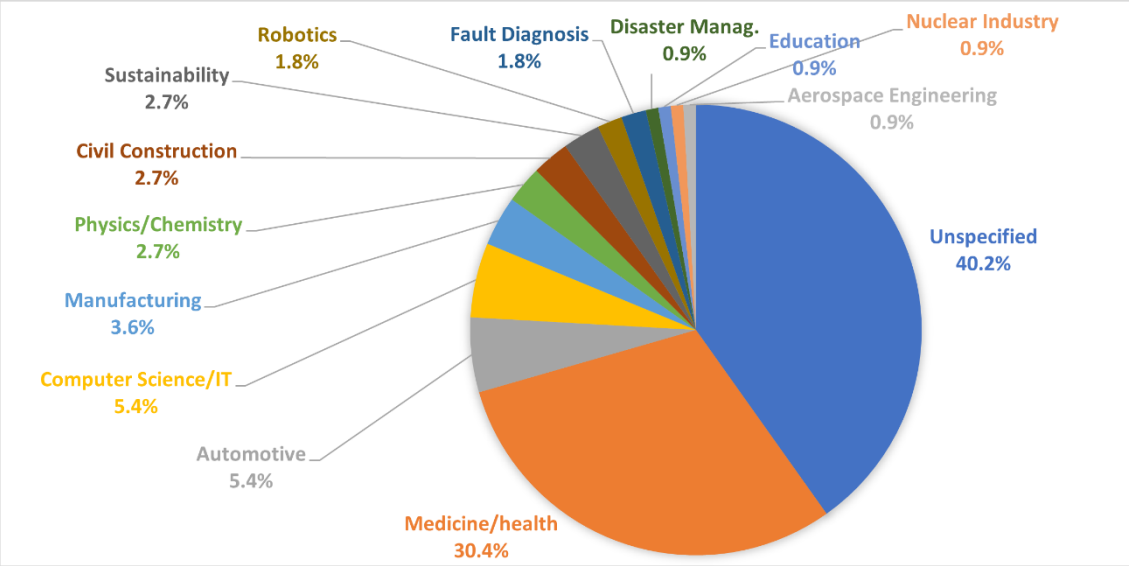
The other techniques - RNNs, AEs, DRL, GNNs and GANs - were used approximately the same number of times. Also, in a small number of articles, a hybrid approach was employed. Hybrid deep learning models are typically composed of multiple (two or more) deep basic learning models, where the basic model is a discriminative or generative deep learning model (SARKER, 2021). Dikshit and Pradhan (2021) used a hybrid DL model, which combined the CNN approach with an LSTM architecture, to predict meteorological drought at different geographic coordinates. It proved to be an effective approach for that end, especially with the added explainability.

Therefore, to answer RQ2, CNNs are the most common DL technique used with XAI.

2.5 RQ3: MAIN AREAS OF APPLICATION

XAI models can be applied to DL in many different areas, such as medicine/health, robotics, engineering, sustainability, among others. In the case of this review, these areas are represented in Figure 5, below.

Figure 5 - Areas in which XAI has been applied to DL.



Source: (THE AUTHORS, 2022).

The articles analyzed in this review show a significant number of medical applications (30.4%); so, the application of XAI in the medical field seems promising. The fact that explanations are especially important in that field, as an error in a medical system is a potential danger to human lives, is probably the main reason for that. Specifically in the medical/health field, most studies applied XAI in cardiology, for diagnosis of heart diseases. Jo *et al*. (2021a) used an explainable DL model to detect and classify arrhythmia.

Furthermore, “Unspecified” is the category with the highest percentage (40.2%), which simply means that the XAI and DL techniques were not applied to advance a specific field. That is the case with those studies that only aim to improve the transparency of DL models in general, such as the study by La Gatta *et al*. (2021), in which they propose a novel XAI technique, CASTLE, for local explanations. The technique can be used in any area, and it was tested on six different datasets containing data from various fields.

Two other areas which are significantly represented in the studies are Automotive Engineering and Computer Science (5.4%). Research on autonomous vehicles has been carried out by authors who used DL and XAI for object detection and recognition in autonomous driving (LI *et al*., 2020). This represents only one application among many in the automotive industry. When XAI and DL are applied to Computer Science, that means they contribute to solving a problem within that area. In order to clarify and provide an example, they have been used in malware detection (IADAROLA *et al*., 2021).

The techniques have also been applied to a great variety of areas besides those already mentioned, like physics, chemistry, manufacturing, sustainability, robotics and others.

Moreover, there are many tasks in which XAI and DL algorithms can used. These tasks

can be image classification, object recognition and detection, among others. In image classification, for example, XAI can be useful in case the objects represented in the images are deformed. When that happens, we cannot rely solely on a black-box DL model decision, as it might not use adequate features for its classification. That is why adding explainability with the aid of XAI is helpful in that kind of situation, since it explains exactly what features the model considered for its decision.

While conducting the review, we found that the tasks that benefited from XAI in DL the most were image classification and object detection / recognition. Also, images were the most common type of data analyzed. Shi *et al*. (2021) developed an explainable DL model to perform automated detection of Geographic Atrophy (GA) presence or absence from OCT volume scans and to provide interpretability by demonstrating which regions of which B-scans show GA. Stavelin *et al*. (2021) aimed to detect fish in images recorded under water and provide insight into the internal workings of the algorithm - this an example of an object detection task. Moreover, Li *et al*. (2020) used DL and XAI for object detection and recognition in autonomous driving.

Thus, to answer RQ3, XAI has contributed to DL in many different areas, especially in medicine / health, and with various types of data and tasks, mostly image classification and object detection.

2.6 RQ4: PERFORMANCE METRICS FOR XAI TECHNIQUES

While numerous explanation techniques have been explored, there is a need for evaluations to assess the quality of explanation techniques to determine whether and to what extent the explainability achieves the defined objective, as well as compare available explanation techniques and suggest the best explanation for a specific task (ZHOU *et al*., 2021).

While conducting this review, we noticed most authors used performance metrics for DL techniques; however, they are not included in the scope of this article. As for metrics to assess the performance of XAI techniques, many authors did not use them, or did not specify them. Specifically, only 42 of the 108 studies employed XAI metrics, which is approximately 38.9%, less than half. Besides, we have not found this as a future research suggestion in the articles analyzed, even though assessing explanations is an important step to ensure their quality.

There are qualitative metrics, which are subjective, and quantitative metrics, which are objective, for XAI (ZHOU *et al*., 2021).

Although some articles have used qualitative metrics, there were more quantitative ones.

Various quantitative metrics have been employed, such as: Maximum Sensitivity (MS) (MEISTER *et al*., 2021a; MEISTER *et al*., 2021b; KAKOGEORGIOU; KARANTZALOS, 2021), Infidelity (MEISTER *et al*., 2021a; MEISTER *et al*., 2021b), Most Relevant First (MoRF) (KAKOGEORGIOU; KARANTZALOS, 2021), and Correctness (KENNY *et al*., 2021; ANTWARG *et al*., 2021; CHEN; LEE, 2020).

As for qualitative metrics, an example is having specialists in the field analyze the explanations, as in the study by Venugopal *et al*. (2020), where the heatmaps were analyzed by a radiologist with more than 8 years' experience in chest imaging. This was a commonly used metric in the studies analyzed (SCHOONDERWOERD *et al*., 2021; JO *et al*., 2021a; SABOL *et al*., 2020), specifically as regards medical applications. The other ones depended on the opinion of the users, who rated, for example, the explanations' interpretability (MORADI; SAMWALD, 2021; WEITZ *et al*., 2021; LEE; WAGSTAFF, 2020).

A few authors implemented various XAI metrics in their studies, such as Kakogeorgiou and Karantzalos (2021), who used most of the quantitative metrics mentioned above to evaluate the performance of ten XAI techniques, and found Occlusion, Lime and Grad-CAM techniques were the most interpretable and less sensitive, since they presented the lowest Max-Sensitivity and AUC-MoRF scores.

Overall, the majority of the authors, about 61.1%, did not employ or specify any XAI performance metrics. Among those that used them, there was not a specific metric that stood out as the most used; however, MS and Infidelity are among the most popular. We consider it an important step for future studies to implement more XAI metrics.

2.6 RQ5: LIMITATIONS AND FUTURE RESEARCH SUGGESTIONS

A common limitation claimed by the authors was that the dataset or the model might not have been adequate, or the reduced size of the dataset (AL HAMMADI *et al*., 2021; LEE *et al*., 2021; HU; MELLO, GAŠEVIĆ., 2021).

Besides, some authors reported that their research could have benefited from having more diversified sources of data. Shi *et al*. (2021) aimed to detect Geographic Atrophy (GA), and the authors believe their results could have been better if volume scans without GA and those with GA were balanced. Others report few age groups, few factors considered and overall small dataset size as limitations (AL HAMMADI *et al*., 2021; LEE *et al*., 2021; BEHL *et al*., 2021). This is an example of the aforementioned limitation of the small sample size, given that

more sources of data are able to provide a higher number of samples and thus increase the size of the dataset.

Also considered limitations were the number and interpretability of explanations provided by the XAI algorithm. A significant number of authors considered that the explanations were not sufficient to clearly understand how the model worked to reach its decision or were difficult to understand by end users, since they are not familiar with XAI explanations (YEGANEJOU; DICK; MILLER, 2020; ISLAM *et al*., 2020; WEITZ *et al*., 2021). Thus, it must be admitted that interpretative models can provide false assurances of comprehensibility. As Páez (2019) states, "the task ahead for XAI is thus to fulfill the double desiderata of finding the right fit between the interpretative and the black box model, and to design interpretative models and devices that are easily understood by the intended users".

Computational cost was also considered an issue in some articles, despite the good results achieved by the explainable DL models. For instance, Kakogeorgiou and Karantzalos (2021) found that LIME and Occlusion had a high computational cost.

Furthermore, as usual, limitations are included in the conclusion section of the articles, but most articles have their limitations stated as future research. That is, most authors have decided to suggest future research based on the limitations they found, so that the authors of future articles will know what they should do to improve their studies and have more accurate results. Therefore, the aforementioned common limitations are also what needs to be considered in future research.

As mentioned, one of the limitations was the difficulty of understanding explanations by end users, and that can be improved in future studies. The generation of explanations within the bounds of the conceptual and linguistic framework of human behavior could greatly improve the transparency and explainability of AI systems towards end-users (WEITZ *et al*., 2021). Selvaraju *et al*. (2020) believe that a true AI system should not only be intelligent, but also be able to reason about its beliefs and actions for humans to trust and use it.

Therefore, common limitations are: reduced size of the dataset, inadequate model / dataset, limited amount of data sources, not enough explanations or not easily understandable by end users, and high computational cost. Suggestions for future research include overcoming these limitations.

2.7 CONCLUSION

This systematic literature review aimed to analyze the contribution of XAI techniques in explaining decisions of DL algorithms. Specifically, to summarize the studies which answer the RQ's, that is, the most used DL and XAI techniques, as well as XAI performance metrics, common limitations, and what is suggested for the future.

We found that XAI has been applied to DL in many fields, such as civil construction, computer science, sustainability, automotive, among others. However, one definitely stands out, which is the medical field. This field in particular can highly benefit from XAI techniques to explain black-box models' predictions, as a prediction based on irrelevant features is possibly a danger to human lives.

As for the XAI techniques, even though SHAP and LIME were employed the most, other techniques, like LRP and Grad-CAM, were not far behind. Authors such as Lee, Jeon and Lee (2021), and López-Cabrera *et al*. (2021) reported similar findings.

Regarding DL techniques, CNNs are very popular in the literature, as they are the DL technique most frequently used with XAI.

Many studies did not employ metrics for assessing XAI techniques. Besides, we have not found this as a future research suggestion in the articles analyzed, but we believe it should be explored in the future. Also, quantitative metrics seem to be used more than qualitative metrics.

Some of the most frequently stated limitations in the studies are: small size of the dataset, inadequate model or dataset, limited amount of data sources, not enough explanations or explanations that end users do not easily interpret, and high computational cost. Suggestions for future research include overcoming these limitations.

Despite that, it is important to highlight the overall good results obtained from the application of DL and XAI techniques together.

For future research, we suggest increasing the time period, a limitation in this review, and exploring some other research questions, such as comparing the effectiveness between XAI applied to DL models and to other models; investigating if metrics will be employed more frequently, and the frequency of the introduction of new XAI techniques, as well as comparing the results of new techniques with state-of-the-art techniques.

We believe the advancement of the explainable DL field and its solutions will result in a fairer, safer and more confident use of DL across society.

REFERENCES

ADADI, A.; BERRADA, M. Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). **IEEE Access**, n. 6, p. 52138–52160, 2018.

AGARWAL, S.; GUDI, R.; SAXENA, P. Application of Computer Vision Techniques for Segregation of Plastic Waste based on Resin Identification Code. **ArXiv Preprint**, ArXiv:1804.08199, p. 1–9, 2020. http://arxiv.org/abs/2011.07747.

AL HAMMADI, A. Y. *et al.* Explainable artificial intelligence to evaluate industrial internal security using EEG signals in IoT framework. **Ad Hoc Networks**, v. 123, 102641, 2021.

ANCONA, M.; CEOLINI, E.; ÖZTIRELI, C.; GROSS, M. **Towards better understanding of gradient-based attribution methods for Deep Neural Networks**. Proceedings of the 6th International Conference on Learning Representations, ICLR 2018, p. 1–16. Vancouver, BC, Canada, 2018.

ANTWARG, L.; MILLER, R. M.; SHAPIRA, B.; ROKACH, L. Explaining anomalies detected by autoencoders using Shapley Additive Explanations. **Expert Systems with Applications**, v. 186, 115736, 2021.

BEHL, S. *et al*. Twitter for disaster relief through sentiment analysis for COVID-19 and natural hazard crises. **International Journal of Disaster Risk Reduction**, v. 55, 102101, 2021.

CHEN, H-Y.; LEE, C-H. Vibration Signals Analysis by Explainable Artificial Intelligence (XAI) Approach: Application on Bearing Faults Diagnosis. **IEEE Access**, v. 8, p. 134246-134256, 2020

DAS, A.; RAD, P. **Opportunities and Challenges in Explainable Artificial Intelligence (XAI): A Survey**. ArXiv, abs/2006.11371, 2020.

DIKSHIT, A.; PRADHAN, B. Interpretable and explainable AI (XAI) model for spatial drought prediction. **Science of The Total Environment**, v. 801, 149797, 2021.

HASSANZADEH, T.; ESSAM, D.; SARKER, R. EvoDCNN: An evolutionary deep convolutional neural network for image classification. **Neurocomputing**, v. 488, p. 271–283, 2022.

HU, Y.; MELLO, R. F.; GAŠEVIĆ, D. Automatic analysis of cognitive presence in online discussions: An approach using deep learning and explainable artificial intelligence. **Computers and Education: Artificial Intelligence**, v. 2, 100037, 2021.

IADAROLA, G.; MARTINELLI, F.; MERCALDO, F.; SANTONE, A. Towards an interpretable deep learning model for mobile malware detection and family identification. **Computers & Security**, v. 105, 102198, 2021.

ISLAM, M. *et al*. Enabling Explainable Fusion in Deep Learning With Fuzzy Integral Neural Networks. **IEEE Transactions on Fuzzy Systems**, v. 28, n. 7, p. 1291-1300, 2020.

JANIESCH, C.; ZSCHECH, P.; HEINRICH, K. Machine learning and deep learning. **Electronic Markets**, v. 31, n. 3, p. 685–695, 2021.

JIA, X.; REN, L.; CAI, J. Clinical implementation of AI technologies will require interpretable AI models. **Medical Physics**, v. 47, n. 1, p. 1–4, 2020.

JO, Y-Y. *et al*. Explainable artificial intelligence to detect atrial fibrillation using electrocardiogram. **International Journal of Cardiology**, v. 328, p. 104-110, 2021.

KENNY, E. M.; FORD, C.; QUINN, M.; KEANE, M. T. Explaining black-box classifiers using post-hoc explanations-by-example: The effect of explanations and error-rates in XAI user studies. **Artificial Intelligence**, v. 294, 103459, 2021.

KINKEAD, M.; MILLAR, S.; MCLAUGHLIN, N.; O'KANE, P. Towards Explainable CNNs for Android Malware Detection. **Procedia Computer Science**, v. 184, p. 959-965, 2021.

LA GATTA, V.; MOSCATO, V.; POSTIGLIONE, M.; SPERLÌ, G. CASTLE: Cluster-aided space transformation for local explanations. **Expert Systems with Applications**, v. 179, 115045, p. 2-12, 2021.

LECUN, Y. *et al*. Backpropagation Applied to Handwritten Zip Code Recognition. **Neural Computation**, v. 1, n. 4, p. 541–551, 1989.

LECUN, Y.; BENGIO, Y.; HINTON, G. Deep learning. **Nature**, v. 521, 7553, p. 436–444, 2015.

LEE, J. H.; WAGSTAFF, K. L. Visualizing image content to explain novel image discovery. **Data Mining and Knowledge Discovery**, v. 34, p. 1777-1804, 2020.

LI, Y. *et al*. A Deep Learning-Based Hybrid Framework for Object Detection and Recognition in Autonomous Driving. **IEEE Access**, v. 8, p. 194228-194239, 2020.

LINARDATOS, P.; PAPASTEFANOPOULOS, V.; KOTSIANTIS, S. Explainable AI: A Review of Machine Learning Interpretability Methods. **Entropy**, v. 23, n. 1, p. 18, 2020.

LÓPEZ-CABRERA, J. D. *et al*. Current limitations to identify covid-19 using artificial intelligence with chest x-ray imaging (part ii). The shortcut learning problem. **Health and Technology**, v. 11, n. 6, p. 1331–1345, 2021.

LØVER, J.; GJÆRUM, V. B.; LEKKAS, A. M. Explainable AI methods on a deep reinforcement learning agent for automatic docking. **IFAC-PapersOnLine**, v. 54, n. 16, p. 146-152, 2021.

MEISTER, S.; WERMES, M.; STÜVE, J.; GROVES, R. M. Cross-evaluation of a parallel operating SVM – CNN classifier for reliable internal decision-making processes in composite inspection. **Journal of Manufacturing Systems**, v. 60, p. 620-639, 2021a.

MEISTER, S.; WERMES, M.; STÜVE, J.; GROVES, R. M. Investigations on Explainable Artificial Intelligence methods for the deep learning classification of fibre layup defect in the automated composite manufacturing. **Composites Part B: Engineering**, v. 224, 109160,

2021b.

MORADI, M.; SAMWALD, M. Post-hoc explanation of black-box classifiers using confident itemsets. **Expert Systems with Applications**, v. 165, 113941, p. 1-14, 2021.

MURDOCH, W. J. *et al*. Definitions, methods, and applications in interpretable machine learning. **Proceedings of the National Academy of Sciences**, v. 116, n. 44, p. 22071–22080, 29 out. 2019.

NEVES, I. *et al*. Interpretable heartbeat classification using local model-agnostic explanations on ECGs. **Computers in Biology and Medicine**, v. 133, 104393, 2021.

PÁEZ, A. The Pragmatic Turn in Explainable Artificial Intelligence (XAI). **Minds and Machines**, v. 29, n. 3, p. 441–459, 2019.

PASQUADIBISCEGLIE, V.; CASTELLANO, G.; APPICE, A.; MALERBA, D. **FOX: a neuro-Fuzzy model for process Outcome prediction and eXplanation**. Proceedings of the 3rd International Conference on Process Mining, ICPM 2021, p. 112-119. Eindhoven University of Technology, De Zaale, Eindhoven, Netherlands, 2021.

RAI, A. Explainable AI: from black box to glass box. **Journal of the Academy of Marketing Science**, v. 48, n. 1, p. 137–141, 2020.

RAS, G.; XIE, N.; VAN GERVEN, M.; DORAN, D. Explainable Deep Learning: A Field Guide for the Uninitiated. **Journal of Artificial Intelligence Research**, v. 73, p. 329–397, 25 jan. 2022.

RIBEIRO, M. T.; SINGH, S.; GUESTRIN, C. **"Why Should I Trust You?": Explaining the Predictions of Any Classifier**. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco, California, USA: ACM, 2016.

RUDIN, C. Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead. **Nature Machine Intelligence**, v. 1, n. 5, p. 206–215, 2019.

SABOL, P. *et al*. Explainable classifier for improving the accountability in decision-making for colorectal cancer diagnosis from histopathological images. **Journal of Biomedical Informatics**, v. 109, 103523, 2020.

SAMEK, W.; WIEGAND, T.; MÜLLER, K.-R. **Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models**. 2017.

SARKER, I. H. Deep Learning: A Comprehensive Overview on Techniques, Taxonomy, Applications and Research Directions. **SN Computer Science**, v. 2, n. 6, p. 420, 2021.

SCHÖNHOF, R. *et al*. Feature visualization within an automated design assessment leveraging explainable artificial intelligence methods. **Procedia CIRP**, v. 100, p. 331-336, 2021.

SCHOONDERWOERD, T. A. J.; JORRITSMA, W.; NEERINCX, M. A.; VAN DEN BOSCH, K. Human-centered XAI: Developing design patterns for explanations of clinical decision support systems. **International Journal of Human-Computer Studies**, v. 154, 102684, 2021.

SELVARAJU, R. R. *et al*. **Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization**. Proceedings of the IEEE International Conference on Computer Vision, ICCV 2017, p. 618-626. Venice, Italy, 2017.

SHAMSHIRBAND, S.; RABCZUK, T.; CHAU, K.-W. A Survey of Deep Learning Techniques: Application in Wind and Solar Energy Resources. **IEEE Access**, v. 7, p. 164650–164666, 2019.

SHI, X. *et al*. Improving Interpretability in Machine Diagnosis: Detection of Geographic Atrophy in OCT Scans. **Ophthalmology Science**, v. 1, n. 3, 100038, 2021.

SIDDAWAY, A. P.; WOOD, A. M.; HEDGES, L. V. How to Do a Systematic Review: A Best Practice Guide for Conducting and Reporting Narrative Reviews, Meta-Analyses, and Meta-Syntheses. **Annual Review of Psychology**, v. 70, n. 1, p. 747–770, 2019.

STAVELIN, H.; RASHEED, A.; SAN, O.; HESTNES, A. J. Applying object detection to marine data and exploring explainability of a fully convolutional neural network using principal component analysis. **Ecological Informatics**, v. 62, 101269, 2021.

VOULODIMOS, A.; DOULAMIS, N.; DOULAMIS, A.; PROTOPAPADAKIS, E. Deep Learning for Computer Vision: A Brief Review. **Computational Intelligence and Neuroscience**, v. 2018, p. 1–13, 2018.

WEITZ, K. *et al*. "Let me explain!": exploring the potential of virtual agents in explainable AI interaction design. **Journal on Multimodal User Interfaces**, v. 15, n. 7, p. 87-98, 2021.

YEGANEJOU, M.; DICK, S.; MILLER, J. Interpretable Deep Convolutional Fuzzy Classifier. **IEEE Transactions on Fuzzy Systems**, v. 28, n. 7, p. 1407-1419, 2020.

ZHOU, J.; GANDOMI, A. H.; CHEN, F.; HOLZINGER, A. Evaluating the Quality of Machine Learning Explanations: A Survey on Methods and Metrics. **Electronics**, v. 10, n. 5, p. 593, 4 mar. 2021.

3

ARTICLE 2: EXPLAINABLE ARTIFICIAL INTELLIGENCE APPLIED TO DEEP LEARNING FOR PLASTICS SORTING

This section refers to the second article, entitled Explainable Artificial Intelligence Applied to Deep Learning for Plastics Sorting.

3.1 INTRODUCTION

Plastic is a versatile and strong material, and its production has increased nearly tenfold since 1950, for a variety of applications (GEYER et al., 2017). Development of synthetic polymers, used to make plastics such as polyethylene, polypropylenes, polyesters and polyamides, has revolutionized the types of containers for products, the types of materials for packaging and other products made of plastic (NKWACHUKWU et al., 2013). However, most of these polymers are not biodegradable; therefore, when they are used and discarded, they can become waste and pollute the environment for a very long time, which may be harmful to human health and the environment (SARDON; DOVE, 2018).

Recycling plastics is a way to reduce environmental impact and resource depletion, as it can decrease energy and material usage per unit of output and so yield improved eco-efficiency (HOPEWELL; DVORAK; KOSIOR, 2009). There are various ways of recycling plastics and the ease of recycling depends on the type of polymer, product and package design. For example, rigid containers that consist of a single polymer are simpler and more economic to recycle than multi-layer and multi-component packages (HOPEWELL; DVORAK; KOSIOR, 2009).

An important step in recycling plastics is sorting them. This separation process is necessary because the presence of even a small quantity of a different type of plastic may decrease the quality of the whole batch (DODBIBA; FUJITA, 2004). Some techniques for sorting plastic materials are as follows: wet separating techniques, such as flotation of plastics (SHIBATA et al., 1996); dry techniques, such as near-infrared spectroscopic analysis or x-rays (WILLIAMS; NORRIS, 1987); and sorting by melting, which can only be used to separate two plastic types at a time (RUJ et al., 2015). There is also the possibility of manually separating the plastics; however, it is not as efficient as automated sorting and tends to take longer. As Jimoh, Ajayi and Ayilara (2014) state, automated sorting systems are necessary in order to achieve high throughput and accuracy.

Besides the aforementioned techniques for separating plastics, there are artificial intelligence and computer vision algorithms for that task. For example, Jimoh, Ajayi and Ayilara (2014) used a fuzzy model to classify images of plastic materials into their respective categories. Additionally, Meeradevi, Raju and Vigneshkumaran (2020) classified plastic bottle images with Convolutional Neural Networks (CNNs), which constitute a type of Deep Learning (DL) model.

DL is a subset of Machine Learning (ML) which typically consists of Artificial Neural Networks (ANN) with more than one hidden layer, organized in deeply nested network architectures (JANIESCH; ZSCHECH; HEINRICH, 2021). DL algorithms have been used in a great variety of fields, such as healthcare, manufacturing, autonomous robots and vehicles, cyber-security, sustainability, as well as with many types of data, for example, image processing, classification and detection, speech and audio processing, among others. They can also provide solutions for sorting plastics in an automated way, as studies such as Meeradevi, Raju and Vigneshkumaran (2020) have shown.

However, DL algorithms are considered black boxes, which means they are not reliable, since their decisions are not transparent (LINARDATOS; PAPASTEFANOPOULOS; KOTSIANTIS, 2020). We cannot understand in which characteristics DL models base their decisions, and therefore, they may consider irrelevant characteristics, thus making them unreliable. To tackle this challenge, explainable AI (XAI) methods can provide human interpretable explanations to better understand machine learning black-box decisions (KAKOGEORGIOU; KARANTZALOS, 2021).

Explainable artificial intelligence (XAI) is a recent research topic which can be a solution to overcome constraints in classical DL methods. XAI techniques applied to DL

algorithms are useful for analyzing individual predictions and discovering the characteristics that contributed the most to making these predictions (RIBEIRO et al., 2016).

In the literature, studies that have used DL combined with XAI techniques for image classification are, for example: Shi et al. (2021) developed an explainable DL model to perform automated detection of Geographic Atrophy (GA) presence or absence from OCT volume scans and to provide interpretability by demonstrating which regions of which B-scans show GA. Furthermore, Saleem, Shahid and Raza (2021) applied CAM, Grad-CAM and Gradient Backpropagation to identify brain tumor in Magnetic Resonance Imaging (MRI) images. Mandeep and Malhi (2020) incorporated XAI in CNNs to classify synthetic aperture radar images.

Specifically in the case of plastics, XAI can help identify the characteristics in plastic packages that caused the DL algorithm to classify them into their respective plastic type. If the package is wrongly attributed to a certain category, XAI techniques provide insights into the reason for the error. Some studies have employed ML and DL techniques for classifying plastics. For example, Tandler et al. (1995) were some of the first researchers to propose an automated way, with knowledge-based expert systems and rule-building ex-pert systems, for something related to sorting plastics, specifically for analyzing the products of catalyzed thermal cracking of polyethylene terephthalate (PET). More recently, Meeradevi, Raju and Vigneshkumaran (2020) used CNNs to classify and sort different types of plastic bottles. However, in a systematic literature review conducted prior to this study, studies on XAI applied to DL for plastics sorting could not be found.

Therefore, in this study, we aim to fill this gap in the literature by applying DL and XAI techniques to plastics sorting and evaluating them quantitatively. In particular, we use CNNs to classify a dataset of plastic images into their respective categories, compare their performance, and, in order to better understand the decisions of the CNNs, we employ Guided Grad-CAM and Integrated Gradients, which are XAI techniques. Finally, to evaluate the XAI techniques quantitatively, we utilize the metrics Max-Sensitivity (MS) and Infidelity.

We expect to contribute with an automated way to the recycling of plastics, especially to the essential step of sorting them.

3.2 MATERIALS AND METHODS

In this section, we describe the type of research, as well as the XAI and DL techniques and quantitative evaluation metrics employed in this study.

This research is categorized as quantitative, since DL and XAI techniques, as well as quantitative metrics, are applied to a dataset of images of plastics for image classification. It is also an applied research because it is an investigation that aims to acquire new knowledge and is directed towards a specific, practical aim, and resolves a practical problem (OECD, 2015).

The dataset defined for the classification task is “Plastic Recycling Codes” (YA, 2020), which is available on the “Kaggle” (kaggle.com) website and open to the public. This dataset contains 8 folders, each one with images of a specific type of plastic, except the last one, in which the images do not actually contain any kind of plastic. Specifically, the first folder contains 121 images of polyethylene (PET) packages; the second, 76 images of high-density polyethylene (HDPE); the third, 24 images of polyvinylchloride (PVC); the fourth, 90 images of low-density polyethylene (LDPE); the fifth, 192 images of polypropylene (PP); the sixth, 39 images of polystyrene (PS); the seventh, 64 of other plastics; and the last one has 79 images of non-plastics. Table 1 contains this information.

Table 1 - Description of the quantity and type of images in each data folder

Folder No.	**Type**	**Quantity**
1	Polyethylene (PET)	121
2	High-density polyethylene (HDPE)	76
3	Polyvinylchloride (PVC)	24
4	Low-density polyethylene (LDPE)	90
5	Polypropylene (PP)	192
6	Polystyrene (PS)	39
7	Other plastics	64
8	Non-plastics	79
TOTAL		**685**

Source: (THE AUTHORS, 2022).

Figure 1 shows some examples of the data images, specifically one from each category.

Figure 1 - Examples of the plastic images. Top row, from left to right: PET, HDPE, PVC, LDPE. Bottom row, from left to right: PP, PS, other plastics, non-plastic.

Source: Extracted from Kaggle (2022).

The images were resized to a suitable size of 224 x 224. Also, ratio of the training, validation, and test sets is 80%, 10%, and 10%, respectively. The model was trained with 100 epochs.

It is important to explain that we have previously conducted a systematic literature review on XAI applied to DL and, in this review, we found the most used XAI and DL techniques. Based on that, in this study, we use CNNs, as they are certainly the most commonly employed technique in the literature. Specifically, we used VGG19, VGG16, ResNet50, and ResNet152, as these architectures are commonly used in the literature for image classification. As regards XAI techniques, we use Guided Grad-CAM and Integrated Gradients (IG), since IG and Grad-CAM are used quite often in studies with DL and XAI techniques, and they are suitable for image data. Finally, we have chosen two common XAI metrics: Max-Sensitivity (MS) and Infidelity, as they are the most used metrics in studies to assess XAI techniques' performance.

ResNet50 (HE et al., 2015) is a variation of the ResNet (residual neural network) architecture with 50 deep layers pre-trained on at least one million images from the ImageNet database. ResNet152 is another type of ResNet and it contains 152 layers, as the name indicates (HE et al., 2015).

The VGG network is a pre-trained CNN model proposed by Simonyan and Zisserman (2015). It was trained on the ImageNet ILSVRC dataset with 1.3 million images. VGG19 is a variant of the VGG architecture and it has 19 layers, while VGG16 has 16 layers (SIMONYAN; ZISSERMAN, 2015).

Grad-CAM (SELVARAJU et al., 2017) assigns values of importance to each neuron in CNNs using the gradient information that flows into the last convolutional layer, for a particular

decision. This method computes the gradients of the output $f_c(x)$ w.r.t. feature map activations A^k of a given layer. The gradients are then averaged for each channel k (along width W and height H) to obtain the importance weights. This is described in equation 1:

$$a_k^c = \frac{1}{H \cdot W} \sum_i^W \sum_j^H \frac{\partial f_c}{\partial A_{ij}^k}(x) \tag{1}$$

Guided Grad-CAM is a combination of Guided Backpropagation, which visualizes fine-grained details in the image, and Grad-CAM, which is class-discriminative and localizes relevant image regions (SELVARAJU et al., 2017). It multiplies the Grad-CAM outcomes with Guided Backpropagation (GBP) values (MEISTER et al., 2021). Since Grad-CAM has been used in many studies, and GBP has also been used quite frequently, we used their combination, Guided Grad-CAM, which combines the best of both techniques.

Integrated Gradients (IG) (SUNDARAJAN et al., 2017) combines the Implementation Invariance of Gradients along with the Sensitivity of techniques like LRP or DeepLift. It is the integral of the gradients along the straight-line path from a baseline $x' = (x_1', \ldots, x_D')$ to the input $x = (x_1, \ldots, x_D)$:

$$\phi_{IG}^d(f_{c'}x) = (x_d - x'_d) \times \int_0^1 \frac{\partial f_c(\tilde{x})}{\partial \tilde{x_d}}_{x=x'+a(x-x')} da \;\forall\, d \in \{1, \ldots, D\}, \tag{2}$$

Regarding the metrics, Max-Sensitivity (MS) (YEH et al., 2019) describes the sensitivity of an XAI algorithm for infinitesimally small modifications in an input data set. Afterwards, this measure is determined based on the normalized difference of the results produced by an XAI method. For calculating the difference, a modified and a reference dataset are considered (MEISTER et al., 2021). This technique has an upper limit and indicates the MS of XAI techniques to disturbances. This metric can be defined as shown in equation 3:

$$SENS_{MAX}(\phi_f, I_m, R_m) = max\|\phi_f(R_m) - \phi_f(I_m)\| \\ , with\ \|R_m - I_m\| \leq r, \tag{3}$$

where r is a customizable value range and the absolute value $\| \cdots \|$ is calculated using the L_2 norm (YEH et al., 2019).

The Infidelity metric (YEH et al., 2019) expresses the correlation between an XAI evaluation and the corresponding CNN model. It describes the relevance of a single input pixel in relation to the CNN response (MEISTER et al., 2021). Equation 4 describes this metric:

$$INFD(\phi_f, I_m, R_m) = \mathbb{E}_{R_{m\sim\mu}}\big[(R_m^T \phi_f(I_m) - \\ (f(I_m) - f(I_m - R_m)))^2\big], \tag{4}$$

where the respective reference R_m is formulated as in equation 5:

$$R_m = I_m - X_0, \quad (5)$$

where X_0 is a random variable which has the probability distribution μ. Also, the expectation value is approximated through a Monte-Carlo calculation (MEISTER et al., 2021).

The implementation was conducted using Python and the PyTorch library, on Google Colab. In addition, we used a HP laptop with Intel Core I7, 8 GB RAM, and Windows 11.

3.3 RESULTS

The following subsections, 3.3.1 and 3.3.2, present the results of the implementation of the DL and XAI techniques, as well as the metrics.

3.3.1 CNN Performance

In order to assess and compare the different CNN architectures employed, we used the metrics accuracy, precision, recall, and F1-score, as shown in Table 2.

Table 2 - Performance of the CNNs.

	ResNet50	**ResNet152**	**VGG16**	**VGG19**
Accuracy (%)	76	79	74	60
Precision (%)	76	81	71	59
Recall (%)	76	79	74	60
F1-Score (%)	73	79	71	58

Source: (THE AUTHORS, 2022).

It is noticeable that VGG19 had the lowest values for all the metrics, which means it might not be the most suitable architecture for this task. On the other hand, ResNet152 obtained the highest scores, being the only one with more than 80% precision. Also, both ResNet50 and 152 had higher scores than VGG16 and 19, which suggests that ResNet is the one that performs better.

This is in line with Mascarenhas and Agarwal (2021), who compared the VGG16, VGG19, and ResNet50 architectures based on their accuracy, and concluded that ResNet50 was the best. Many studies also consider ResNet a better architecture, as it is deeper.

Moreover, it is possible that the accuracy would increase if the number of epochs was increased, as was the case in the study by Ikechukwu et al. (2021), in which the overall accuracy was similar to this study with 100 epochs, the same as in this study, but increased significantly when they changed to 300 epochs.

Furthermore, it is important to mention that the precision for each class significantly varied, which could be caused by the unbalanced data.

3.3.2 Performance and Results of XAI Techniques

In order to quantify the reliability of the employed XAI methods, we utilized Max-Sensitivity and Infidelity. Table 3 shows the results of the quantitative metrics for each XAI technique combined with each CNN.

Table 3 - Quantitative XAI metrics (lower scores indicate higher performance).

	ResNet 50 + IG	**ResNet50 + Grad-CAM**	**ResNet152 + IG**	**ResNet152 + Grad-CAM**	**VGG16 + IG**	**VGG16 + Grad-CAM**	**VGG19 + IG**	**VGG19 + Grad-CAM**
Max-Sensitivity	0.75	0.32	0.76	0.39	0.46	0.27	0.48	0.28
Infidelity	0.04	0.01	0.04	0.68	0.06	0.08	0.07	0.15

Source: (THE AUTHORS, 2022).

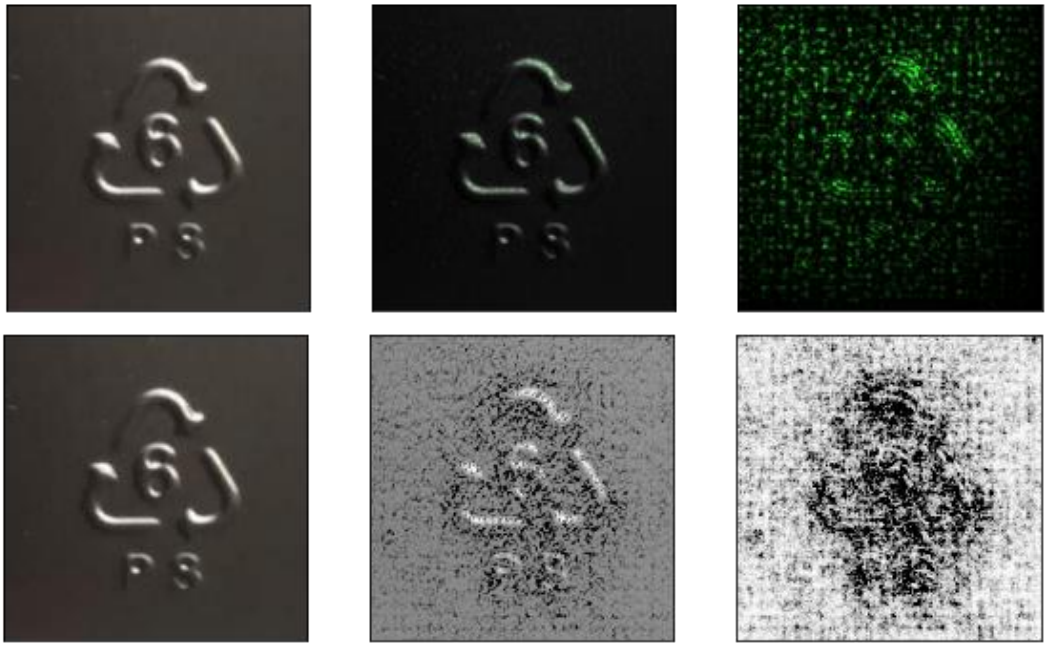
Regarding the metric MS, Guided Grad-CAM achieved the lowest scores, when compared to IG, for each architecture and even among all the architectures.

However, the Infidelity is slightly higher for Guided Grad-CAM with all architectures, and much higher with ResNet152, where it reached 0.68. Guided Grad-CAM used to interpret the decisions of ResNet152 also obtained the highest MS compared to the same technique used with other networks. This might mean that Guided Grad-CAM is less effective for explaining the decisions of ResNet152.

On the other hand, the other Infidelity values were very close and low, therefore there is no significant difference in infidelity between the two XAI techniques with the other CNNs. In addition, MS and Infidelity values were within the range of what is usually found in the literature, although not many studies have used these metrics. For example, Kakogeorgiou and Karantzalos (2021), Meister et al. (2021), and Sahatova and Balabaeva (2022) employed these metrics and obtained similar results.

Figure 2 below shows an example of Guided Grad-CAM and IG applied on the same PS image, classified with ResNet152.

Figure 2 - The same PS image with Guided Grad-CAM (top) and IG (bottom).

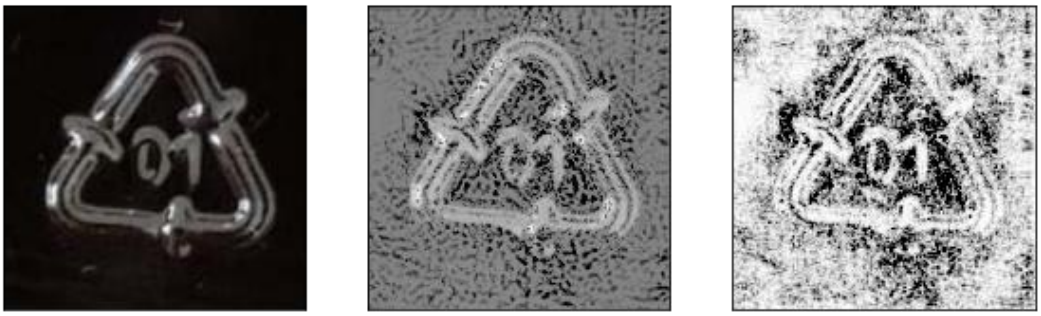


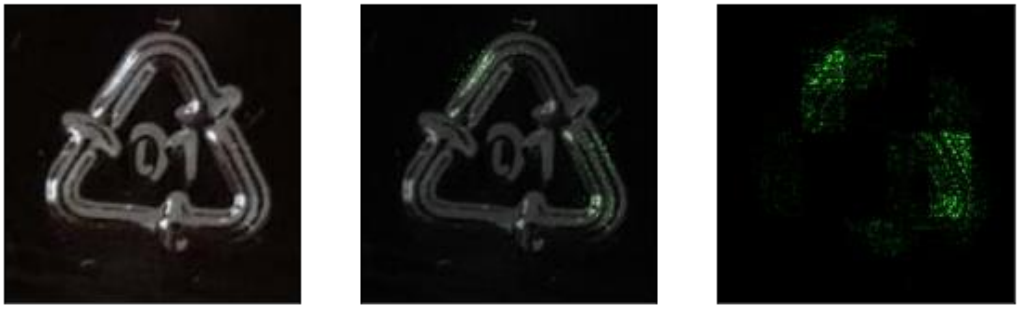
Source: (THE AUTHORS, 2022).

It is possible to see that Guided Grad-CAM, on the top row, showed that the relevant characteristics were mostly around the identification of the type of plastic, which would mean that the model used inadequate characteristics to make its decisions. However, Guided Grad-CAM infidelity for this instance was very high, which means the explanation is likely not accurate. As opposed to that, IG focused mainly on the inside of the triangle and the letters that identified the resin, and the infidelity was much lower. This happened with other images as well, thus suggesting Guided Grad-CAM is less effective than IG to explain the predictions of ResNet152.

In some cases, the images were misclassified, and the XAI techniques showed that it was because the model considered inadequate regions of the image for classification. Figure 3 below shows an example of this.

Figure 3 - A misclassified image by VGG16 with IG (top) and Guided Grad-CAM (bottom).





Source: (THE AUTHORS, 2022).

In this particular instance, we can see that the model considered irrelevant regions, as IG highlights the regions around the number and triangle, and Guided Grad-CAM, with less noise, highlights some regions of and around the triangle. Thus, the importance of XAI techniques lies in their ability to show why an image was classified into a certain class and explain the reason for misclassifications.

In addition, some misclassifications happened despite having relevant regions considered. For example, Figure 4 shows an image in which the letters and number corresponding to the plastic were highlighted, but the classification was still wrong.

Figure 4 - A misclassified image by VGG16 with IG (top) and Guided Grad-CAM (bottom).

Source: (THE AUTHORS, 2022).

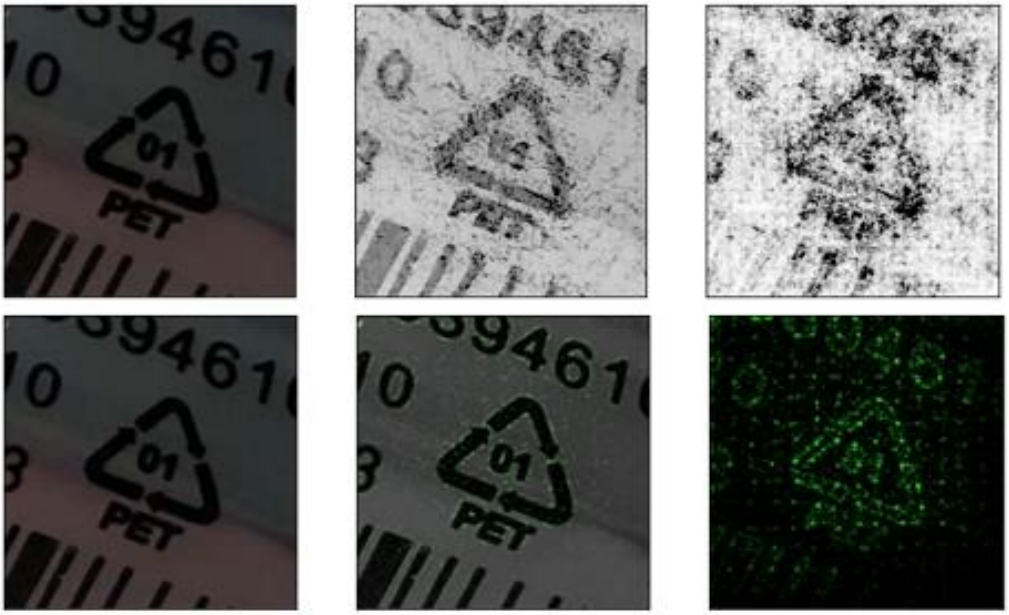
In fact, both XAI methods show the correct regions as relevant, namely, the initials referring to the type of plastic, and the numbers, although the latter are less emphasized. However, it was still classified as LDPE. This might be because the image is flipped horizontally, and the CNN considered the edge of the triangle as relevant, which might be confused with a “4”, or it is also possible that the letter “T” has been considered an “L”.

Moreover, it was common for the networks to wrongly classify images into the class of non-plastics when they considered regions that had many irrelevant characteristics, such as text or numbers that were not the plastic identification. This is probably due to the fact that many images of non-plastics contained a lot of text and other features. Figure 5 shows some examples of non-plastic images, and Figure 6 is an example of misclassification likely due to irrelevant information in the picture.

Figure 5 - Examples of original images of non-plastics.

Source: (THE AUTHORS, 2022).

Figure 6 - PET image classified by ResNet152 as a non-plastic and explained using IG (top) and Guided Grad-CAM (bottom).

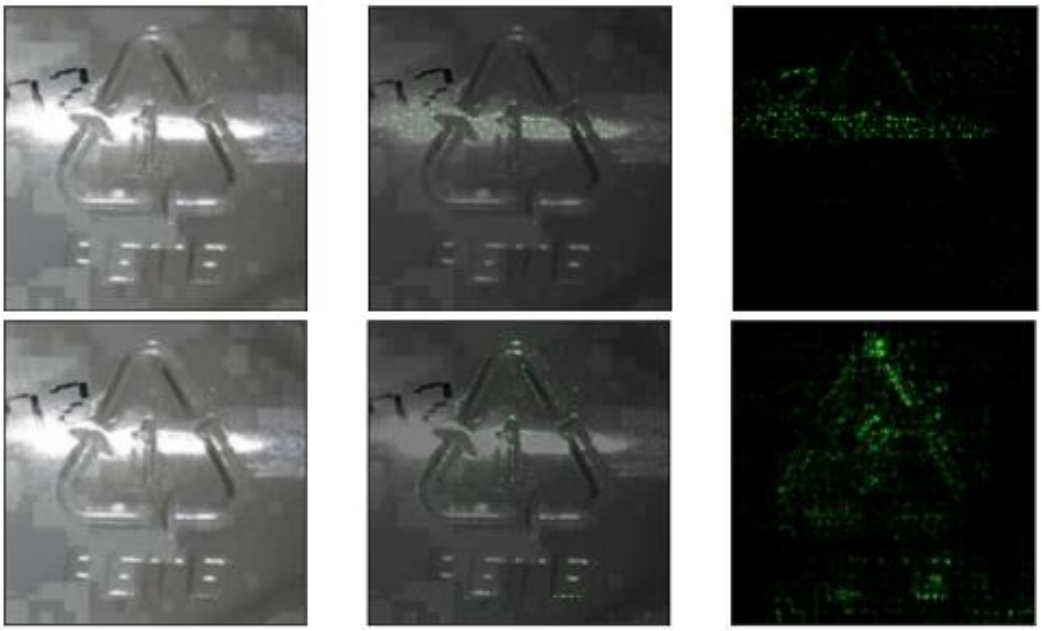


Source: (THE AUTHORS, 2022).

Both XAI techniques show that the model considered the extra information around the triangle, which was irrelevant and probably caused the mistake, since images of non-plastics also tend to contain more information. This misclassification related to extra information occurred in all the models, but even more frequently in ResNet152.

Another case of misclassification is when it was not very clear which type of plastic the image contained, or there were some interfering elements, such as the color, low contrast or reflected light. Figure 7 contains an image which was misclassified by ResNet152 but correctly classified by VGG19.

Figure 7 - An image misclassified by ResNet152 (top) and correctly classified by VGG19 (bottom) with Guided Grad-CAM.



Source: (THE AUTHORS, 2022).

It is clear that, as Grad-CAM shows, the first model considered the reflected light as important for the classification, and that is probably the reason for the misclassification; meanwhile, the model represented on the bottom made a correct decision because it considered more relevant regions of the image.

Overall, Guided Grad-CAM appears to be a better XAI technique than IG, as it has less noise and seems to focus more on specific pixels and regions, which makes it clearer which parts of the images were considered important for the models' decisions. This is noticeable in the figures above, as well as in most explanations. Furthermore, the metric MS indicates that Guided Grad-CAM is less sensitive and thus more robust. The only exception seems to be when this technique is used to explain ResNet152, as mentioned above.

3.4 DISCUSSION

According to our findings, ResNet152 had the highest accuracy, precision, recall, and F1-score. Both ResNet models performed better than VGG. This finding is in line with Mascarenhas and

Agarwal (2021), who compared the VGG16, VGG19, and ResNet50 architectures based on their accuracy, and found that ResNet50 performed better. Other studies, such as Triyadi, Bustamam and Anki (2022) have also found ResNet to perform better when used for image classification. The reason for that is likely because ResNet is deeper than VGG, and depth is important in image classification (SIMONYAN; ZISSERMAN, 2015).

Regarding the XAI techniques, MS indicates that Guided Grad-CAM produces better explanations than IG, and that is also noticeable by observing the explanations in the images. However, Infidelity indicates that this technique was worse for interpreting the decisions of ResNet152. The relevant regions are clearly identified by Guided Grad-CAM in most images classified by VGG, but less frequently in images classified by ResNet152. Besides, Guided Grad-CAM identifies specific regions more clearly and with less noise than IG, thus making its explanations more interpretable. Kakogeorgiou and Karantzalos (2021) obtained different results for one of the datasets they used, but similar ones for the other, that is, in their study, MS was higher for Guided Grad-CAM when applied to BigEarthNet, but slightly lower when applied to SEN12MS.

In addition, some common reasons for misclassification were irrelevant extra information, low contrast, reflected light and other types of changes and deformation of the plastic. Once again, this shows DL models are unreliable and XAI explanations are very useful to understand them.

Also regarding XAI metrics, MS and Infidelity were similar to the values that are usually found in the literature, although not many studies have used these metrics. Kakogeorgiou and Karantzalos (2021), and Sahatova and Balabaeva (2022) employed these metrics and obtained similar results. Moreover, many studies have employed Smooth IG instead of traditional IG, and obtained lower infidelity and sensitivity. One example is the study by Kakogeorgiou and Karantzalos (2021). Yeh et al. (2019) dedicate an entire section to this explanation, and conclude that smoothing explanations indeed reduces sensitivity and infidelity. Thus, introducing smooth IG and other smoothed techniques could be a suggestion to reduce the MS and Infidelity of the models in this study.

We believe that the results would be different, probably better, if the dataset was bigger and more balanced. The low number of images, particularly from specific classes, is a limitation that might have influenced the results. A suggestion to overcome that limitation in the future is data augmentation.

3.5 CONCLUSIONS

In this study, we aimed to employ CNNs to classify plastic images into the respective type of plastic, and to explain the CNNs decisions by using XAI techniques, as well as metrics to assess the explanations.

We implemented VGG19, VGG16, ResNet50, and ResNet152 and found that ResNet152 had the best performance. Also, both ResNet architectures performed better than VGG. The reason for that result might be that ResNet is a deeper network. However, the performance was similar among the CNNs, only VGG19 had a lower performance. All these models commonly produce good results in image classification and are suitable for the task.

Regarding the XAI techniques, we found that, overall, Guided Grad-CAM produces better explanations, more interpretable, with less noise, and obtained lower MS. Neverthe-less, this technique had worse results when interpreting ResNet152, reaching a significantly higher Infidelity. Therefore, the quality of XAI explanations of each method might depend on the CNN, and it is helpful to test different XAI techniques for different CNN architectures.

As for XAI metrics, MS and Infidelity were similar to the values that are usually found in the literature, although not many studies have used these metrics. Additionally, an interesting way to reduce MS and Infidelity values would be to apply Smooth IG instead of traditional IG, and even other smoothed techniques, as Yeh et al. (2019) explain.

A limitation in this study is the dataset, which is small and unbalanced. We would probably achieve better results if the dataset contained more images and each class had approximately the same number of images. Therefore, a suggestion for future studies is to use a larger dataset or to use data augmentation.

In addition, the next step should be implementing these techniques on real data from plastic recycling industries, and analyzing whether it is feasible to apply these techniques on large-scale data.

Another suggestion that contributes to further research is to compare more XAI techniques and use other metrics, as well as to increase the number of epochs in the training phase.

In general, we conclude that XAI is useful to understand DL models, to visualize and understand which characteristics they consider more relevant for their predictions, and it makes them more transparent and reliable.

REFERENCES

DODBIBA, G.; FUJITA, T. Progress in Separating Plastic Materials for Recycling. **Physical Separation in Science and Engineering**, v. 13, no. 3–4, p. 165–182, 2004.

GEYER, R.; JAMBECK, J.R.; LAW, K.L. Production, use, and fate of all plastics ever made. **Science Advances**, v. 3, no. 7, 2017.

HE, K.; ZHANG, X.; REN, S.; SUN, J. Deep Residual Learning for Image Recognition. In: **Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition** (CVPR), 2016, p. 770-778. doi: 10.1109/CVPR.2016.90.

HOPEWELL, J.; DVORAK, R.; KOSIOR, E. Plastics recycling: challenges and opportunities. **Philosophical transactions of the Royal Society of London**. Series B, Biological sciences, v. 364, no. 1526, p. 2115–2126, 2009.

IKECHUKWU, A.V.; MURALI, S.; DEEPU, R.; SHIVAMURTHY, R.C. ResNet-50 vs VGG-19 vs training from scratch: A comparative analysis of the segmentation and classification of Pneumonia from chest X-ray images. **Global Transitions Proceedings**, v. 2, no. 2, p. 375-381, 2021. https://doi.org/10.1016/j.gltp.2021.08.027

JIMOH, K.O.; AJAYI, A.O.; AYILARA, O.A. Intelligent Model for Manual Sorting of Plastic Wastes. **International Journal of Computer Applications**, v. 101, no. 7, p. 20-26, 2014.

KAKOGEORGIOU, I.; KARANTZALOS, K. Evaluating explainable artificial intelligence methods for multi-label deep learning classification tasks in remote sensing. **International Journal of Applied Earth Observation and Geoinformation**, v. 103, no. 102520, 2021.

LINARDATOS, P.; PAPASTEFANOPOULOS, V.; KOTSIANTIS, S. Explainable AI: A Review of Machine Learning Interpretability Methods. **Entropy**, v. 23, no. 1: 18, 2021.

MANDEEP, H.S.; MALHI, A. Deep learning-based explainable target classification for synthetic aperture radar images. In: **Proceedings of the 2020 13th International Conference on Human System Interaction**, HSI 2020, p. 34-39, IEEE Computer Society. https://doi.org/10.1109/HSI49210.2020.9142658

MASCARENHAS, S.; AGARWAL, M. A comparison between VGG16, VGG19 and ResNet50 architecture frameworks for Image Classification. In: **Proceedings of the 2021 International Conference on Disruptive Technologies for Multi-Disciplinary Research and Applications**, CENTCON, 2021, p. 96-99. doi: 10.1109/CENTCON52345.2021.9687944.

MEERADEVI; RAJU, S.M.; VIGNESHKUMARAN. Automatic Plastic Waste Segregation And Sorting Using Deep Learning Model. **International Journal of Scientific & Technology Research**, v. 9, no. 2, 2020.

MEISTER, S.; WERMES, M.; STÜVE, J.; GROVES, R. M. Investigations on Explainable Artificial Intelligence methods for the deep learning classification of fibre layup defect in the automated composite manufacturing. **Composites Part B: Engineering**, v. 224, 109160, 2021

NKWACHUKWU, O.I.; CHIMA, C.H.; IKENNA, A.O.; ALBERT, L. Focus on potential environmental issues on plastic world towards a sustainable plastic recycling in developing countries. **International Journal of Industrial Chemistry**, v. 4, no. 34, 2013.

OECD. Frascati Manual 2015: Guidelines for Collecting and Reporting Data on Research and Experimental Development, The Measurement of Scientific, Technological and Innovation Activities. OECD Publishing, Paris, 2015. DOI: http://dx.doi.org/10.1787/9789264239012-en

RIBEIRO, M.T.; SINGH, S.; GUESTRIN, C. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In: **Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining**, 2016, p. 1135–1144.

RUJ, B.; PANDEY, V.; JASH, P.; SRIVASTAVA, V. Sorting of plastic waste for effective recycling. **International Journal of Applied Sciences and Engineering Research**, v. 4, no. 4, 2015.

SAHATOVA, K.; BALABAEVA, K. An Overview and Comparison of XAI Methods for Object Detection in Computer Tomography. **Procedia Computer Science**, v. 212, p. 209-219, 2022.

SALEEM, H.; SHAHID, A.R.; RAZA, B. Visual interpretability in 3D brain tumor segmentation network. **Computers in Biology and Medicine**, v. 133, 104410, 2021.

SARDON, H.; DOVE, A.P. Plastics recycling with a difference. **Science**, v. 360, no. 6387, 2018.

SELVARAJU, R.R.; COGSWELL, M.; DAS, A.; VEDANTAM, R.; PARIKH, D.; BATRA, D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. **International Journal of Computer Vision**, v. 128, p. 336–359, 2020.

SHI, X.; KEENAN, T.; CHEN, Q.; SILVA, T.; THAVIKULWAT, A.; BROADHEAD, G.; BHANDARI, S.; CUKRAS, C.; CHEW, E.; LU, Z. Improving Interpretability in Machine Diagnosis: Detection of Geographic Atrophy in Optical Coherence Tomography Scans. **Ophthalmology Science**, v. 1, no. 3, 100038, 2021.

SHIBATA, J; MATSUMOTO, S.; YAMAMOTO, H.; KUSAKA, E.; PRADIP, P. Flotation separation of plastics using selective depressants. **International Journal of Mineral Processing**, v. 48, no. 3, p. 127-134, 1996.

SIMONYAN, K.; ZISSERMAN, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. In: **Proceedings of the 3rd International Conference on Learning Representations**, ICLR 2015, San Diego, California, USA.

TANDLER, P.J.; BUTCHER, J.A.; TAO, H.; HARRINGTON, P. B. Analysis of plastic recycling products by expert systems. **Analytica Chimica Acta**, v. 312, p. 231-244, 1995.

WILLIAMS, P.; NORRIS, K. H. Near-infrared technology in the agricultural and food industries. **American Association of Cereal Chemists**, St. Paul, Minn., USA: American Association of Cereal Chemists, 1987.

YA, P. Plastic Recycling Codes. Dataset, 2020. Available at: <https://www.kaggle.com/piaoya/plastic-recycling-codes>.

YEH, C.-K.; HSIEH, C.-Y.; SUGGALA, A.S.; INOUYE, D.I.; RAVIKUMAR, P. On the (In)fidelity and Sensitivity of Explanations. In: **Proceedings of the 33rd Conference on Neural Information Processing Systems** (NeurIPS 2019), Vancouver, Canada, arXiv:1901.09392, 2019.

4

CONCLUSION

The aim of this dissertation was to verify which DL and XAI techniques, as well as XAI metrics, are most used together, in the literature, and also to apply DL and XAI techniques to classify plastic images and thus contribute to plastics sorting. In order to do that, the first step was to conduct a SLR on XAI applied to DL. Based on the results of the SLR, DL and XAI techniques and metrics were chosen and employed to classify a dataset of plastic images into their respective types. This was divided in two studies: one containing the SLR, and the other referring to the implementation of the algorithms.

The SLR provided some insights into the application of XAI in DL, which constitute the first results of this work and served as a basis for the applied study. It was possible to observe that XAI has been applied to DL in many different fields, but one definitely stands out, which is the medical field.

Furthermore, CNNs are very popular in the literature, as they are the DL technique most frequently used in conjunction with XAI.

As for the XAI techniques, the conclusion is not as clear. Even though SHAP was employed the most, other techniques, such as LIME, LRP and Grad-CAM, were not far behind.

Moving on to the metrics for assessing XAI techniques, overall, a great variety of XAI performance metrics was used, and many authors did not employ or specify them. However, MS and Infidelity seem to be among the most used.

Based on the findings of the SLR, CNNs were implemented, specifically ResNet50, ResNet152, VGG19 and VGG16, some of the most used architectures. Performance metrics

showed that ResNet152 achieved a better performance than the other models. Since it is common that deeper networks perform better, that might be the reason for that result.

In addition, XAI techniques Integrated Gradients and Guided Grad-CAM were applied to explain the CNNs' decisions. The findings indicate that Guided Grad-CAM produces better explanations, in general, as it obtained the best MS and Infidelity scores, and also has less noise and focus on more specific regions of the images. There was an exception, though, when it was used with ResNet152, as the metrics were higher and the explanations not as interpretable.

Regarding the XAI metrics, the obtained values were within the values found in the literature.

Some suggestions for future SLRs are increasing the time period, a limitation in this review, and exploring some other research questions, such as comparing the effectiveness between XAI applied to DL models and to other models; investigating if metrics will be employed more frequently, and the frequency of the introduction of new XAI techniques, as well as comparing the results of new techniques with state-of-the-art techniques. For practical studies, some suggestions are: increasing the size of the dataset and using a balanced one, either by using a larger dataset or data augmentation, comparing more XAI techniques and using other metrics, as well as increasing the number of epochs in the training phase. These suggestions would probably help overcome the limitations found in this research, such as small dataset size, unbalanced dataset, and short time period considered for the SLR.

This dissertation is expected to contribute to further advance research on XAI applied to DL, which is a recent topic and can certainly benefit from more research. It also proposes an automated way to sort plastics to be recycled, which can be useful for the plastic recycling industry.

REFERENCES

ADADI, A.; BERRADA, M. Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). In: **IEEE Access**, v. 6, p. 52138-52160, 2018.

ANCONA, M.; CEOLINI, E.; ÖZTIRELI, C.; GROSS, M. Towards better understanding of Gradient-based Attribution Methods for Deep Neural Networks. In: **Proceedings of the International Conference on Learning Representations**, 6th, 2018, Vancouver, Canada.

AYRE, D. Technology advancing polymers and polymer composites towards sustainability: A review. **Current Opinion in Green and Sustainable Chemistry**, v. 13, p. 108-112, 2018.

BACH, S.; BINDER, A.; MONTAVON, G.; KLAUSCHEN, F.; MÜLLER, K.-R.; SAMEK, W. On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation. **PLOS ONE**, v. 10, no. 7, 2015.

BEHL, S.; RAO, A.; AGGARWAL, S.; CHADHA, S.; PANNU, H.S. Twitter for Disaster Relief Through Sentiment Analysis for COVID-19 and Natural Hazard Crises. **International Journal of Disaster Risk Reduction**, v. 55, 102101, 2021.

BOBULSKI, J.; KUBANEK, M. Deep Learning for Plastic Waste Classification System. **Applied Computational Intelligence and Soft Computing**, v. 2021, p, 1-7, 2021.

DIKSHIT, A.; PRADHAN, B. Interpretable and explainable AI (XAI) model for spatial drought prediction. **Science of the Total Environment**, v. 801, 149797, 2021.

DODBIBA, G.; FUJITA, T. Progress in Separating Plastic Materials for Recycling. **Physical Separation in Science and Engineering**, v. 13, no. 3–4, p. 165–182, 2004.

GEYER, R.; JAMBECK, J.R.; LAW, K.L. Production, use, and fate of all plastics ever made. **Science Advances**, v. 3, no. 7, 2017.

GILPIN, L.H.; BAU, D.; YUAN, B. Z.; BAJWA, A.; SPECTER, M.; KAGAL, L. Explaining Explanations: An Overview of Interpretability of Machine Learning. In: **2018 IEEE 5th International Conference on Data Science and Advanced Analytics** (DSAA), Turin, Italy, 2018, pp. 80-89.

GRUBER, F.; GRÄHLERT, W.; WOLLMANN, P.; KASKEL, S. Classification of Black Plastics Waste Using Fluorescence Imaging and Machine Learning. **Recycling**, v. 4, no. 40, 2019.

HE, K.; ZHANG, X.; REN, S.; SUN, J. Deep Residual Learning for Image Recognition. In: **Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition** (CVPR), 2016, p. 770-778. doi: 10.1109/CVPR.2016.90.

HENRIKSEN, M. L.; KARLSEN, C. B.; KLARSKOV, P.; HINGE, M. Plastic classification via in-line hyperspectral camera analysis and unsupervised machine learning. **Vibrational Spectroscopy**, v. 118, no. 103329, 2022.

HOLDEN, E.; LINNERUD, K. The sustainable development area: satisfying basic needs and safeguarding ecological sustainability. **Sustainable Development**, v. 15, no. 3, p. 174-187, 2007.

HOOKER, S.; ERHAN, D.; KINDERMANS, P.-J.; KIM, B. Evaluating feature importance estimates. In: arXiv preprint, arXiv:1806.10758, 2018.

HOPEWELL, J.; DVORAK, R.; KOSIOR, E. Plastics recycling: challenges and opportunities. **Philosophical transactions of the Royal Society of London**. Series B, Biological sciences, v. 364, no. 1526, p. 2115–2126, 2009.

HU, Y.; MELLO, R.F.; GAŠEVIĆ, D. Automatic analysis of cognitive presence in online discussions: An approach using deep learning and explainable artificial intelligence. **Computers and Education: Artificial Intelligence**, v. 2, 100037, 2021.

HUBER, T.; WEITZ, K.; ANDRÉ, E.; AMIR, O. Local and global explanations of agent behavior: Integrating strategy summaries with saliency maps. **Artificial Intelligence**, v. 301, 103571, 2021.

HUMMELS, H.; ARGYROU, A. Planetary demands: Redefining sustainable development and sustainable entrepreneurship. **Journal of Cleaner Production**, v. 278, no. 123804, 2021.

IKECHUKWU, A.V.; MURALI, S.; DEEPU, R.; SHIVAMURTHY, R.C. ResNet-50 vs VGG-19 vs training from scratch: A comparative analysis of the segmentation and classification of Pneumonia from chest X-ray images. **Global Transitions Proceedings**, v. 2, no. 2, p. 375-381, 2021. https://doi.org/10.1016/j.gltp.2021.08.027

JANIESCH, C.; ZSCHECH, P.; HEINRICH, K. Machine learning and deep learning. **Electron Markets**, v. 31, p. 685–695, 2021.

JIMOH, K.O.; AJAYI, A.O.; AYILARA, O.A. Intelligent Model for Manual Sorting of Plastic Wastes. **International Journal of Computer Applications**, v. 101, no. 7, p. 20-26, 2014.

KAKOGEORGIOU, I.; KARANTZALOS, K. Evaluating explainable artificial intelligence methods for multi-label deep learning classification tasks in remote sensing. **International Journal of Applied Earth Observation and Geoinformation**, v. 103, no. 102520, 2021.

KASEVA, M.E.; GUPTA, S.K. Recycling — an environmentally friendly and income generating activity towards sustainable solid waste management. Case study — Dar es Salaam City, Tanzania. **Resources, Conservation and Recycling,** v. 17, no. 4, p. 299-309, 1996.

KINKEAD, M.; MILLAR, S.; MCLAUGHLIN, N.; O'KANE, P. Towards Explainable CNNs for Android Malware Detection. **Procedia Computer Science**, v. 184, p. 959-965, 2021.

LA GATTA, V.; MOSCATO, V.; POSTIGLIONE, M.; SPERLÌ, G. CASTLE: Cluster-aided space transformation for local explanations. **Expert Systems with Applications**, v. 179, 115045, 2021.

LECUN, Y.; BENGIO, Y.; HINTON, G. Deep learning. **Nature**, v. 521, no. 7553, p. 436–444, 2015.

LINARDATOS, P.; PAPASTEFANOPOULOS, V.; KOTSIANTIS, S. Explainable AI: A Review of Machine Learning Interpretability Methods. **Entropy**, v. 23, no. 1: 18, 2021.

LØVER, J.; GJÆRUM, V.B.; LEKKAS, A.M. Explainable AI methods on a deep reinforcement learning agent for automatic docking. In: **Proceedings of the IFAC Conference on Control Applications in Marine Systems, Robotics, and Vehicles CAMS**, 13th, 2021, Oldenburg, Germany. IFAC, v. 54, no. 16, 2021, p. 146-152.

MANDEEP, H.S.; MALHI, A. Deep learning-based explainable target classification for synthetic aperture radar images. In: **Proceedings of the 2020 13th International Conference on Human System Interaction**, HSI 2020, p. 34-39, IEEE Computer Society. https://doi.org/10.1109/HSI49210.2020.9142658

MASCARENHAS, S.; AGARWAL, M. A comparison between VGG16, VGG19 and ResNet50 architecture frameworks for Image Classification. In: **Proceedings of the 2021 International Conference on Disruptive Technologies for Multi-Disciplinary Research and Applications**, CENTCON, 2021, p. 96-99. doi: 10.1109/CENTCON52345.2021.9687944.

MEERADEVI; RAJU, S.M.; VIGNESHKUMARAN. Automatic Plastic Waste Segregation And Sorting Using Deep Learning Model. **International Journal of Scientific & Technology Research**, v. 9, no. 2, 2020.

MEISTER, C.; LAZOV, S.; AUGENSTEIN, I.; COTTERELL, R. Is Sparse Attention more Interpretable?. In: **Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing**, v. 2: Short Papers, Online. Association for Computational Linguistics, 2021, p. 122–129.

MEISTER, S.; WERMES, M.; STÜVE, J.; GROVES, R. M. Investigations on Explainable Artificial Intelligence methods for the deep learning classification of fibre layup defect in the automated composite manufacturing. **Composites Part B: Engineering**, v. 224, 109160, 2021b.

MORADI, M.; SAMWALD, M. Post-hoc explanation of black-box classifiers using confident itemsets. **Expert Systems with Applications**, v. 165, 113941, 2021.

MURDOCH, W. J.; SINGH, C.; KUMBIER, K.; ABBASI-ASL, R.; YU, B. Definitions, methods, and applications in interpretable machine learning. In: **Proceedings of the National Academy of Sciences of the United States of America**, v. 116, no. 44, p. 22071–22080, 2019.

NKWACHUKWU, O.I.; CHIMA, C.H.; IKENNA, A.O.; ALBERT, L. Focus on potential environmental issues on plastic world towards a sustainable plastic recycling in developing countries. **International Journal of Industrial Chemistry**, v. 4, no. 34, 2013.

PÁEZ, A. The Pragmatic Turn in Explainable Artificial Intelligence (XAI). **Minds and Machines**, v. 29, no. 3, p. 441-459, 2019.

RAHMAN, M.A.; HOSSAIN, M.S.; SHOWAIL, A.J.; ALRAJEH, N.A.; ALHAMID, M.F. A secure, private, and explainable IoHT framework to support sustainable health monitoring in a smart city. **Sustainable Cities and Society**, v. 72, 103083, 2021.

RAI, A. Explainable AI: from black box to glass box. **Journal of the Academy of Marketing Science**, v. 48, p. 137–141, 2020.

RAS, G.; XIE, N.; VAN GERVEN, M.; DORAN, D. Explainable Deep Learning: A Field Guide for the Uninitiated. **Journal of Artificial Intelligence Research**, v. 73, 2022.

RIBEIRO, M.T.; SINGH, S.; GUESTRIN, C. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In: **Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining**, 2016, p. 1135–1144.

RUDIN, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. **Nature Machine Intelligence**, v. 1, no. 5, p. 206-215, 2019.

RUJ, B.; PANDEY, V.; JASH, P.; SRIVASTAVA, V. Sorting of plastic waste for effective recycling. **International Journal of Applied Sciences and Engineering Research**, v. 4, no. 4, 2015.

SARDON, H.; DOVE, A.P. Plastics recycling with a difference. **Science**, v. 360, no. 6387, 2018.

SCHÖNHOF, R.; WERNER, A.; ELSTNER, J.; ZOPCSAK, B.; AWAD, R.; HUBER, M. Feature visualization within an automated design assessment leveraging explainable artificial intelligence methods. **Procedia CIRP**, v. 100, p. 331-336, 2021.

SCOTT, D.M. A two-color near infrared sensor for sorting recycled plastic waste. **Measurement Science and Technology**, v. 6, no. 2, p. 156-159, 1995.

SELVARAJU, R.R.; COGSWELL, M.; DAS, A.; VEDANTAM, R.; PARIKH, D.; BATRA, D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. **International Journal of Computer Vision**, v. 128, p. 336–359, 2020.

SAHATOVA, K.; BALABAEVA, K. An Overview and Comparison of XAI Methods for Object Detection in Computer Tomography. **Procedia Computer Science**, v. 212, p. 209-219, 2022.

SALEEM, H.; SHAHID, A.R.; RAZA, B. Visual interpretability in 3D brain tumor segmentation network. **Computers in Biology and Medicine**, v. 133, 104410, 2021.

SHAMEEM, K.; CHOUDHARI, K.S.; BANKAPUR, A.; KULKARNI, S.D.; UNNIKRISHNAN, V.K.; GEORGE, S.D.; KARTHA, V.B.; SANTHOSH, C. A hybrid LIBS-Raman system combined with chemometrics: an efficient tool for plastic identification and sorting. **Analytical and bioanalytical chemistry**, v. 409, no. 13, p. 3299–3308, 2017.

SHI, X.; KEENAN, T.; CHEN, Q.; SILVA, T.; THAVIKULWAT, A.; BROADHEAD, G.; BHANDARI, S.; CUKRAS, C.; CHEW, E.; LU, Z. Improving Interpretability in Machine Diagnosis: Detection of Geographic Atrophy in Optical Coherence Tomography Scans. **Ophthalmology Science**, v. 1, no. 3, 100038, 2021.

SHIBATA, J; MATSUMOTO, S.; YAMAMOTO, H.; KUSAKA, E.; PRADIP, P. Flotation separation of plastics using selective depressants. **International Journal of Mineral Processing**, v. 48, no. 3, p. 127-134, 1996.

SIMONYAN, K.; ZISSERMAN, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. In: **Proceedings of the 3rd International Conference on Learning Representations**, ICLR 2015, San Diego, California, USA.

STAVELIN, H.; RASHEED, A.; SAN, O.; HESTNES, A.J. Applying object detection to marine data and exploring explainability of a fully convolutional neural network using principal component analysis. **Ecological Informatics**, v. 62, 101269, 2021

TANDLER, P.J.; BUTCHER, J.A.; TAO, H.; HARRINGTON, P. B. Analysis of plastic recycling products by expert systems. **Analytica Chimica Acta**, v. 312, p. 231-244, 1995.

UITERKAMP, B.J.S.; AZADI, H.; HO, P. Sustainable recycling model: A comparative analysis between India and Tanzania. **Resources, Conservation and Recycling**, v. 55, no. 3, p. 344-355, 2011.

VILLAS, M.; MACEDO-SOARES, T.; RUSSO, G. Bibliographical Research Method for Business Administration Studies: A Model Based on Scientific Journal Ranking. **BAR: Brazilian Administration Review**, v. 5, no. 2, art. 4, p. 139-159, 2008. Available online at: <http://www.anpad.org.br/bar>.

WEITZ, K.; SCHILLER, D.; SCHLAGOWSKI, R; HUBER, T.; ANDRÉ, E. “Let me explain!”: exploring the potential of virtual agents in explainable AI interaction design. Journal on Multimodal User Interfaces, v. 15, p. 87–98, 2021.

WILLIAMS, P.; NORRIS, K. H. Near-infrared technology in the agricultural and food industries. **American Association of Cereal Chemists**, St. Paul, Minn., USA: American Association of Cereal Chemists, 1987.

World Commission on Environment and Development (WCED). **Our Common Future**. WCED–Oxford University Press: Oxford, 1987.

YA, P. Plastic Recycling Codes. Dataset, 2020. Available at: <https://www.kaggle.com/piaoya/plastic-recycling-codes>.

YEH, C.-K.; HSIEH, C.-Y.; SUGGALA, A.S.; INOUYE, D.I.; RAVIKUMAR, P. On the (In)fidelity and Sensitivity of Explanations. In: **Proceedings of the 33rd Conference on Neural Information Processing Systems** (NeurIPS 2019), Vancouver, Canada, arXiv:1901.09392, 2019.

APPENDIX A – Selected Articles for the SLR

#	**Study (Authors, Year)**	**XAI Techniques (RQ1)**	**DL Techniques (RQ2)**	**XAI in DL Applications (RQ3)**	**XAI Metrics (RQ4)**	**XAI in DL Limitations and Future Research (RQ5)**
1	Saleem, Shahid, Raza (2021)	Grad-CAM, Guided BP	DMFNet	Medicine / Health	Deletion metric	Improve segmentation accuracy
2	Quellec *et al*. (2021)	ExplAIn-CAM, ExplAIn	CNN	Medicine / Health	NA	Limited to binary or multilabel classification; pixel-level evaluation relies on incomplete lesion segmentations; ExplAIn would be even more useful for a totally new classification problem in medicine
3	Al Hammadi *et al*. (2021)	SHAP, Permutation Feature Importance	CNN	Medicine / Health	NA	Reduced sample size, few age groups considered
4	Neves *et al*. (2021)	Permutation Sample Importance, LIME, SHAP	CNN	Medicine / Health	Faithfulness, Jaccard Index, Performance decrease	Other types of explanations, such as text-based

5	Schoonderwoerd, Neerincx, van den Bosch (2021)	DoReMi	CNN	Medicine / Health	Specialist's opinion	Further test explanations on their understandability, trust development and decision-making performance; include different types of use cases and users
6	Onchis, Gillich (2021)	LIME, SHAP	MLP	Civil Construction	NA	Re-samplings were necessary
7	Jo *et al*. (2021)	Embedded, Grad-CAM, Guided BP	CNN	Medicine / Health	Specialist's opinion	Use other features of ECG and conduct more practical studies
8	Wu *et al*. (2021)	Embedded	GAN	Medicine / Health	NA	Use a cycle-consistency loss instead of pixel-wise loss; extend GAN to 3D architecture; optimize the structure and components of the technique
9	Lee *et al*. (2021)	LIME	CNN	Medicine / Health	NA	Small dataset, few data sources
10	Liz *et al*. (2021)	Heatmaps	CNN	Medicine / Health	NA	Could use other techniques and increase dataset quality
11	Dikshit, Pradhan (2021)	SHAP	CNN-BiLSTM	Sustainability	NA	Analyze the SHAP plots for long lead time forecasting; examine other additive SHAP properties
12	Løver, Gjærum, Lekkas (2021)	SHAP, LIME	DRL	Robotics	NA	Implement other techniques
13	Kenny *et al*. (2021)	COLE-HP	CNN	Unspecified	Correctness, reasonableness, satisfaction, trust	Use other DL techniques
14	Agarwal, Tamer, Budman (2021)	LRP	AE	Chemistry	NA	Limitations for future practical application: lack of available data,

						longer development time for off-line model calibration
15	Venugopal *et al*. (2020)	Grad-CAM	CNN	Medicine / Health	NA	Features observed and described by a single person; handcrafting of the classification patterns of the activation maps
16	Akula *et al*. (2021)	CX-ToM (Counterfactual Explanations – Theory of Mind), LIME, Grad-CAM, Smooth IG, LRP, TCAV, CEM, CVE	LSTM, CNN	Unspecified	Justified Trust, Explanation Satisfaction	NA
17	Behl *et al*. (2021)	LIME	MLP, CNN	Disaster Management	NA	Limited dataset; classification accuracy depends on the kind of resource asked in the specific disaster
18	Iadarola *et al*. (2021)	Grad-CAM	CNN	Computer Science / IT	Average Euclidean Distance (AED)	Not possible to detect malware not belonging to families in the training dataset
19	Meister *et al*. (2021a)	Smooth IG, Grad-CAM, SHAP	CNN	Manufacturing	MS, Infidelity	Limitations for assessing the model fidelity of SHAP
20	Tsimpouris, Tsakiridis, Theocharis (2021)	Embedded	AE, CNN	Geology	NA	Implement more advanced autoencoder techniques; combine AEs with other techniques; apply to other types of data

21	Hu, Mello, Gašević (2021)	Smooth IG	CNN	Computer Science / IT, Education	NA	Reduced sample size and sources
22	Shi *et al*. (2021)	Embedded	CNN	Medicine / Health	NA	Unbalanced dataset; limited data sources; improve the model with other techniques; apply the method on multimodal imaging
23	Gomez-Fernandez *et al*. (2021)	Saliency Maps	CNN	Nuclear Industry	NA	Cross-discipline studies must be encouraged
24	Stavelin *et al*. (2021)	Colormap	MLP	Sustainability	NA	Used a very complicated network to detect only a single class
25	De Souza *et al*. (2021)	Saliency, Guided BP, Smooth IG, Input x Gradient, DeepLIFT	CNN	Medicine / Health	Cohen-Kappa (CK), Pixel Accuracy (PA), Intersection-over-Union (IoU)	Consider more sophisticated and deeper CNN architectures; assess each layer's importance; apply the method to more datasets
26	Yoo, Kang (2021)	Grad-CAM	CNN	Manufacturing	NA	Use Computer Numerical Control (CNC) machining simulations and interviews with CNC experts; use more XAI techniques
27	Sabol *et al*. (2020)	X-CFCMC	CNN	Medicine / Health	Specialist’s opinion	Not fully automated; include more pathologists from various fields; apply the method on imbalanced data
28	Li, Shi, Hwang (2021)	SHAP, Embedded	CNN	Medicine / Health	NA	Use the method in other applications; apply attention mechanism and transfer learning

29	He, Aouf, Song (2021)	SHAP-CAM	DRL, CNN	Aerospace Engineering	NA	Some explanations don't make sense; use other XAI techniques to improve explanations
30	Meister *et al*. (2021b)	Smooth IG	CNN	Manufacturing	Maximum Sensitivity (MS), Infidelity	Use other DL and XAI techniques; apply the method to different image-based inspection processes
31	Kinkead *et al*. (2021)	LIME, Embedded	CNN	Computer Science / IT	NA	Study which individual filters may be learning features relevant to malware and benign apps
32	Yeom *et al*. (2021)	LRP	CNN	Unspecified	NA	Use heatmaps to elucidate and explain which image features are most strongly affected by pruning
33	Kakogeorgiou, Karantzalos (2021)	Saliency, Input × Gradient, Smooth IG, Guided BP, Grad-CAM, DeepLIFT, Occlusion, LIME	CNN	Unspecified	MS, MoRF, File Size, Computation Time	No high-resolution outputs; LIME and Grad-CAM are not computationally efficient
34	Schönhof *et al*. (2021)	Grad-CAM, LIME, LRP, Sensitivity Analysis	CNN	Unspecified	NA	Only especially relevant regions are highlighted; apply the method to other data
35	Bhakte, Pakkiriswamy, Srinivasan (2021)	SHAP	CNN	Fault Diagnosis	NA	Improve the model; extend the method to processes with multiple, distinct normal operating modes; use other neural networks

36	Zeltner *et al*. (2021)	Continuous logic / Squashing functions	CNN	Unspecified	NA	Combine continuous logic with extreme learning machines (ELM)
37	Shimizu *et al*. (2021)	Embedded	GNN	Unspecified	NA	Use a sensitivity test; include more side information; deal with "over smoothing"
38	La Gatta *et al*. (2021)	Cluster-aided space transformation for local explanations (CASTLE), Anchors	MLP	Unspecified	Computation Time, Coverage, Precision	User cannot control trade-off between precision and coverage; decisions are difficult to understand
39	Moradi, Samwald (2021)	Confident Itemsets Explanation (CIE), LIME, MUSE	MLP, RNN	Unspecified	Fidelity, user's Interpretability, Coverage	Modify CIE to be applied in domain-specific tasks; add a mechanism that shows what modifications are needed to change the outcome
40	Antwarg *et al*. (2021)	SHAP	AE	Unspecified	Correctness, Robustness, Sensitivity, Mean Reciprocal Rank (MRR), experts' feedback	Examine the background set used for the explanation model; use more complicated AEs and more datasets
41	Rahman *et al*. (2021)	AIF360, AIX360, Alibi, Captum, Explainax, Grad-CAM,	CNN, RNN	Medicine / Health, Sustainability	NA	Low accuracy; high loss; lack of labeled datasets; difficulty in obtaining datasets; implement other models

		InterpretML, LIME, SHAP				
42	Jo *et al*. (2021)	Embedded	CNN	Medicine / Health	NA	Use other features of ECG; extend XAI to diagnose other diseases; apply it in clinical practice
43	Barnard (2021)	SHAP	MLP	Physics	NA	NA
44	Díez *et al*. (2020)	Embedded	CNN	Computer Science / IT	NA	Ask users to assess the photos proposed by the model; apply synonymity of images
45	Huber *et al*. (2021)	LRP	DRL	Unspecified	Satisfaction, Spearman Rank Correlation, Structural Similarity (SSim), Pearson Correlation	Using saliency maps on videos instead of static images overwhelms users, potentially missing useful information
46	López *et al*. (2021)	Aspect Discovery for OPinion Summarisation	AE	Unspecified	Unusualness; Significance	Enlarge the database; improve extraction of explicit and implicit aspects
47	Zdravković, Ćirić, Ignjatović (2021)	LIME	RNN	Civil Construction	NA	More sustainable solutions
48	Mandeep, Malhi (2020)	LIME	CNN	Automotive	NA	Limitations in radar technologies restrict image resolutions; use diverse datasets which are not public; incorporate Convolutional AE (CAE)

49	Szandała (2021)	Grad-CAM	CNN	Unspecified	Faithfulness, interpretability and applicability (FIA)	Interpretability tools based on heatmaps do not perform augmentation processes
50	Kamakshi, Gupta, Krishnan (2021)	PACE (Post-hoc Architecture Agnostic Concept Extractor)	CNN	Unspecified	Interpretability, Consistency	NA
51	Diallo, Nakagawa, Tsuchiya (2020)	Smooth IG	CNN	Unspecified	NA	Apply the method to IoT platforms
52	Suzuki *et al*. (2021)	Relative Attributing Propagation (RAP)	GAN	Unspecified	Confidence	Conduct quantitative or large-scale subjective evaluation of the methods; apply the method to other datasets
53	Tan, Khan, Guan (2020)	Locality Guided Neural Network (LGNN)	CNN	Unspecified	Correlation between filters	Incorporate “winner takes all”; tune hyper parameters for the neighborhood function for each layer
54	Yeganejou, Dick, Miller (2020)	LRP, Guided BP, Taylor Decomposition	CNN	Unspecified	NA	Cannot be generalized; make the explanations understandable to the general public; use other classifiers
55	Nascita *et al*. (2021a)	SHAP, DeepLIFT	CNN	Automotive	NA	Investigate trustworthiness of traffic classifiers; compare global explanations from other XAI techniques; apply the method to other traffic analysis tasks

56	Hailemariam *et al*. (2020)	LIME, SHAP	CNN	Medicine / Health	Identity, Stability, Separability	Define quantifiable/ objective metrics for information leakage and explanation misuse; needed expert validation for explanations; apply the method to other datasets; use other XAI techniques
57	Kim, Park (2021)	LIME	Deep Convolutional GAN	Unspecified	NA	Quantify the data
58	Zhang *et al*. (2021)	Embedded	CNN	Unspecified	Part interpretability, Location instability	Decreased classification performance when classifying large number of categories; limited applicability; not suitable to encode textural patterns
59	Le, Kang, Kim (2021)	Grad-CAM	CNN	Unspecified	NA	Apply several approaches in defensive system (e.g, WGAN) to build a robust defend technique
60	Islam *et al*. (2020)	Embedded	Choquet Integral Multilayer Perceptron (ChIMP)	Unspecified	NA	Explore efficient representations; investigate advanced learning algorithms; explore where and when a fusion neuron should be used; make XAI explanations understandable to the general public
61	Kamal *et al*. (2021)	LIME	CNN	Medicine / Health	NA	Apply more interactive XAI
62	Gulum, Trombley, Kantardzic (2021)	Grad-CAM, Saliency Maps	CNN	Unspecified	Lesion Localization, Cascading Randomization, Correctness	Create a more general framework for combination; create a weighted combination of explanation techniques

63	Ahn *et al*. (2021)	Embedded	CNN	Automotive	NA	Design a key feature selection algorithm for finer-grained application-specific traffic classifiers; enable real-time key feature selection
64	Sudhakar *et al*. (2021)	Adaptive Semantic Input Sampling (Ada-SISE), SISE, Extremal Perturbation	CNN	Unspecified	Energy-Based Pointing Game (EBPG), Bbox, Drop and Increase rates, Computational Time	Use important features to analyze model's behavior
65	Malolan, Parekh, Kazi (2020)	LIME, LRP, IG, Guided BP	CNN	Unspecified	NA	NA
66	Diallo, Nakagawa, Tsuchiya (2021)	IG	CNN	Unspecified	NA	Implement the method on the planner component of the MAPE-K feedback loop framework
67	Reza *et al*. (2021)	LIME	CNN	Medicine / Health	NA	Weigh the classes according to their distribution; apply background noise removal; apply attention model to focus on Choroid region
68	Chen, Lee (2020)	Grad-CAM	CNN	Fault Diagnosis	Correctness; verification with transparent models	NA
69	Fan *et al*. (2020)	Customizable Model Interpretation Evaluation (CMIE)	CNN	Unspecified	Feature Information Gain, Feature Sparsity, Feature Completeness,	Use another method, such as Network Dissection, to make feature evaluation more significant

					Interpretation Tree Accuracy, Interpretation Tree Completeness	
70	Islam *et al*. (2021)	Saliency Maps, Effective Gradient (EG), Smooth IG	CNN	Unspecified	NA	Explain the functional composition of the operations; interpret morphological networks with generalized operations
71	Ye, Xia, Yang (2021)	LIME, SHAP	CNN	Medicine / Health	NA	Apply the method to other image datasets and clinical questions
72	Han, Park, Hong (2021)	LRP, SHAP	CNN	Fault Diagnosis	NA	Study the effectiveness of current data in the critical fault condition with relatively large features
73	Li *et al*. (2020)	Randomized Input Sampling for Explanation (RISE)	CNN	Automotive	NA	Improve overall speed of the framework; improve performance of the system; integrate distance prediction
74	Jung, Han, Choi (2021)	LRP, Contrastive LRP (CLRP), Softmax gradient LRP (SGLRP)	CNN, RNN	Unspecified	Maximal patch masking; Pointing game; Deletion	When most of the activations have a negative gradient, it is difficult to find the most important part for prediction; use other algorithms
75	Nascita *et al*. (2021b)	LIME, LRP, SHAP, Smooth IG	CNN	Computer Science / IT	NA	Implement occlusion analysis; investigate trustworthiness, interpretability and robustness of the technique; use more DL techniques; design self-explainable DL classifiers

						and lightweight XAI architectures
76	Wong, McPherson (2021)	Embedded	CNN, MLP	Physics	NA	Apply the method to other applications in the same area; use more sophisticated architectures, loss functions, and hyper-parameter tuning
77	Kuppa, Le-Khac (2020)	Input*Gradient (I*G), LRP, Guided BP, Smooth-GRAD, Grad-CAM, IG	MLP, AE	Computer Science / IT	Consistency, Correctness, Confidence	Study defense mechanisms against the attack proposed; extend the method to compromise the privacy and confidentiality of explainable methods; examine security robustness of other XAI with different neural network architectures
78	Onchis (2020)	LIME, SHAP	MLP	Civil Construction	NA	NA
79	Kim, Bansal (2021)	Attentional Bottleneck	CNN	Automotive	Sparsity	Generate instance level attention maps and; use the maps to improve the performance of the baseline driving model
80	Pianpanit *et al*. (2021)	Saliency Maps, Guided BP, Grad-CAM, DeepLIFT, SHAP	CNN	Medicine / Health	Thresholding, Dice coefficient, Wilcoxon signed-rank test, mean absolute error	Use interpreted feedback for deciding the most suitable model; apply methods to other tasks
81	Jiang, Hewner, Chandola (2021)	Embedded	RNN	Medicine / Health	NA	Make explanations understandable; use other types of data
82	Jain *et al*. (2021)	Grad-CAM	GAN, CNN	Medicine / Health	NA	Apply other data augmentation techniques; optimize the number of

						features; tune hyperparameters
83	Taylor, Shekhar, Taylor (2020)	Response Time (RT)	CNN	Unspecified	NA	Extend method to other dynamic inference models and visual tasks and their controls
84	Wang *et al*. (2021)	Guided BP	Deep Unfolding Super-Resolution Network	Medicine / Health	Feature matching	Feature detection focused on high frequency areas; extend the method to the entire image
85	Dong, Ma, Basu (2021)	Guided BP	CNN	Medicine / Health	NA	NA
86	Kohlbrenner *et al*. (2020)	LRP	CNN	Unspecified	Attribution Localization, Object-centricity	NA
87	Carvalho, Silva (2021)	SHAP	RNN	Law	Perceived quality, perceived value	NA
88	Pasquadibisceglie *et al*. (2021)	Neuro-Fuzzy model for process Outcome prediction and eXplanation (FOX)	CNN	Unspecified	Friedman's test; Nemenyi test	Lack of prescription with the explanation of predictions; does not deal with imbalanced condition; conduct training continuously as new events are logged; explore other encoding mechanisms
89	Gupta *et al*. (2021)	Image Retrieval with Textual Explanations	CNN	Unspecified	T-test	The keypoints are not always visible to the naked eye and it is challenging to zoom in and compare

		(IRTEX)				
90	Patil, Framewala, Kazi (2020)	LRP	MLP	Unspecified	NA	Neglects the importance of feature engineering of the dataset; use the model for high dimensional datasets where it is necessary to eliminate noise
91	Bento *et al*. (2021)	LRP	GAN	Unspecified	NA	NA
92	Arrieta *et al*. (2021)	Potential Memory, Temporal Patterns, Pixel Absence Effect	RNN	Unspecified	NA	Lessen propagation of bias; transform spatially correlated data into sequences; take a close look at the interplay between explainability and epistemic uncertainty
93	Lee, Jeon, Lee (2021)	IG, Smooth IG, Guided BP, Deep Taylor, LRP	CNN	Manufacturing	Satisfaction, Goodness	Combine the results of various visualization techniques
94	Lo, Yin (2021)	I-score, Backward Dropping Algorithm	CNN	Medicine / Health	Deletion	Apply the technique to other types of image datasets
95	Biswas, Barz, Sonntag (2020)	Embedded	CNN	Unspecified	NA	Develop segmentation-based visual explanation techniques and compare with state-of-the-art techniques
96	Yoo *et al*. (2021)	Grad-CAM	AE, CNN	Unspecified	NA	Apply DL using 3D data to CAE simulations, considering aesthetics and manufacturing constraints; develop a 3D generative design technique without using 2D images;

97	Weitz *et al.* (2021)	LIME	CNN	Unspecified	Users' evaluation of whether explanations were sufficient	Use more types of explanations and more human-like; make more precise statements about perceived trust
98	Uddin *et al.* (2021)	LIME	RNN	Medicine / Health	NA	Use a more comprehensive dataset; apply to mental healthcare services to predict mood disorders
99	Selvaraju *et al.* (2020)	Grad-CAM	CNN	Unspecified	NA	NA
100	Agarwal *et al.* (2021)	eXpert AUGmented variables (XAUG), LRP	CNN	Physics	NA	NA
101	Hyeon *et al.* (2021)	Grad-CAM	CNN	Medicine / Health	NA	NA
102	Raihan, Nahid (2021)	SHAP	CNN	Medicine / Health	NA	Build a CAD system; improve feature extraction technique and classifier; make a more in-depth comparison with other algorithms; use more datasets
103	Dong *et al.* (2021)	Region of Evidence (ROE)	CNN	Medicine / Health	Professionals' opinion	Scarcity of data, in quantity and quality
104	Lee, Wagstaff (2020)	DEMUD-VIS	CNN	Unspecified	Utility (users' opinion)	Explore fully labeled datasets, to help identify labeling errors and/or adversarial examples
105	Bautista-Montesano, Bustamante-Bello, Ramirez-	Embedded	DRL	Automotive	NA	Develop a more specific set of rules; increase granularity of the steering angle; test the model in other platforms

	Mendoza (2020)					
106	Cruz *et al*. (2021)	Embedded	DRL	Robotics	Users' opinion	Add reward signals; apply study to a real-world scenario
107	Parra-Ullauri *et al*. (2021)	Event-driven Temporal Models for Explanations (ETeMoX)	DRL	Unspecified	NA	Change formulation; define different time window for handovers query; develop other types of explanations for other systems; make temporal model more flexible
108	Mensa *et al*. (2020)	Embedded	CNN-BiLSTM	Medicine / Health	NA	Apply the model to other domains in the medical field and others

Source: (THE AUTHORS, 2022).