applied sciences

MDPI

Article

An AI-Based Approach for the Automatic Classification and Labeling of Joint Implants Using Registry Taxonomy and Manufacturer Data: A PoC Study with an Application in a Real-World Framework

Tiziana Falcone [1,†,‡], Enrico Ciminello [1,‡], Mascia Masciocchi [1], Alessandro Mentuccia [2], Marco Davide Forlani [2], Martina Barral [2], Martina Zanforlini [2], Duilio Luca Bacocco [3], Eugenio Carrani [1], Fernanda Gellona [4], Valeria Glorioso [4], Elaine Young [5], Timothy Wilton [5], Richard Armstrong [6], Edd Caton [6], Paola Ciccarelli [1] and Marina Torre [1,*,‡]

[1] Italian Registry of Implantable Prostheses, National Centre for Clinical Governance, Italian National Institute of Health, Viale Regina Elena, 299, 00161 Rome, Italy; t.falcone@inail.it (T.F.); enrico.ciminello@iss.it (E.C.); mascia.masciocchi@iss.it (M.M.); eugenio.carrani@outlook.com (E.C.); paola.ciccarelli@iss.it (P.C.)

[2] BVTech, Piazza A. Diaz, 6, 20123 Milan, Italy; alessandro.mentuccia@bvtech.com (A.M.); marco.forlani@bvtech.com (M.D.F.); martina.barral@bvtech.com (M.B.); martina.zanforlini@bvtech.com (M.Z.)

[3] Office for the General Affairs, Italian National Institute of Health, Viale Regina Elena, 299, 00161 Rome, Italy; duilioluca.bacocco@iss.it

[4] Confindustria Dispositivi Medici, Via Marco Burigozzo, 1, 20122 Milan, Italy; f.gellona@meddevicelab.it (F.G.); glorioso@confindustriadm.it (V.G.)

[5] National Joint Registry (NJR), 130 Wood Street, London EC2V 6DL, UK; timothy.wilton@njr.org.uk (T.W.)

[6] NEC Software Solutions, 1st Floor, iMex Centre, 575-599 Maxted Road, Hemel Hempstead HP2 7DX, UK; richard.armstrong@necsws.com (R.A.); edward.caton@necsws.com (E.C.)

* Correspondence: marina.torre@iss.it

† Current affiliation: Department of Technological Innovations and Safety of Plants, Products and Anthropic Settlements, National Institute for Insurance Against Accidents at Work (INAIL), Via Roberto Ferruzzi 38-40, 00143 Rome, Italy.

‡ These authors contributed equally to this work.

Abstract

Joint replacement registries are essential for post-market surveillance; however, medical device (MD) classification dictionaries often rely on manual processes and heterogeneous data sources, limiting efficiency and standardization. The aim of this study was to design and validate, as a proof of concept (PoC), an artificial intelligence (AI) classification system for hip and knee implants, leveraging registry taxonomies and manufacturer technical documents to overcome the above limitations. A PoC study was conducted using a custom-built dataset derived from the Italian Arthroplasty Registry (RIAP) MD-Dictionary, the UK National Joint Registry MD-Classification, and 49 manufacturer documents. A convolutional neural network (CNN) and a transformer-based model (BioBERT) were evaluated. Model performance was assessed on a validation dataset (49 documents) and an external test set (25 documents) using standard metrics and a comparison with manual extraction. BioBERT outperformed the CNN across main evaluation metrics (precision 94.62% vs. 67.18%, recall 88.45% vs. 75.82%, and F1-score 91.43% vs. 71.24%). Processing times were reduced from over 22 h for manual extraction to 4 min for CNN and 27 min for BioBERT. The results show that this AI module-PoC enables the extraction of structured data from technical documents and enhances data completeness and surveillance capabilities by integrating NLP-based automation into MD registries.

Check for updates

Academic Editor: Miguel Alcaraz Baños

Received: 29 July 2026
Revised: 2 October 2026
Accepted: 9 October 2026
Published: 28 October 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

Keywords: artificial intelligence; joint implants; medical device classification; medical device surveillance; natural language processing; registry

1. Introduction

The continuous evolution of implantable medical devices (MDs), particularly those used in joint replacement, highlights the need for a robust and equally fast system to monitor their outcomes. For orthopedic surgeons, the ability to track device performance is not only a regulatory requirement but also a clinical necessity [1]. Postoperative outcomes, long-term follow-up, and decision-making in revision surgeries all depend on complete, reliable, and standardized device data [2].

Implant registries are essential tools for the systematic collection of data on device usage, surgical outcomes, and adverse events, thus supporting post-market surveillance and traceability in compliance with the Medical Device Regulation (EU) 2017/745 [3–6], the early identification of device failures, and quality improvement across healthcare services [7–9]. Registry data, when characterized by high completeness and quality, can also support nationwide analyses based on Hospital Discharge Records (HDRs), thus contributing to the evaluation of temporal trends in surgical procedures of joint replacement as well as supporting forecasting approaches [10,11].

In Italy, the Italian Arthroplasty Registry (RIAP) started data collection in 2007 but was formally established within the wider context of the Italian Implantable Prostheses Registry (RIPI) at the Istituto Superiore di Sanità in 2017 [12]. A key component of RIAP is the MD database, the so-called MD-Dictionary that ensures the standardized identification of implantable devices [13–16]. Dictionaries and the related taxonomies facilitate data interoperability across registries, healthcare institutions, and manufacturers. They also support automated data processing and device-level analytics [13,17,18].

However, dictionaries face limitations such as the reliance on data provided by manufacturers, the lack of automatic updates, and the challenge in integrating heterogeneous data sources. On the other hand, artificial intelligence (AI) applications are being used increasingly in different orthopedic areas [19–25] and the potential of AI to improve traceability and post-market MD surveillance has already been highlighted, for instance, through its successful application in harmonizing safety notices from 16 EU countries [26]. Therefore, to address these issues, particularly in the context of MD registries, an AI module was conceived to automate the classification of joint implants by extracting structured information from a variety of web-based sources (e.g., manufacturer websites and related technical documents).

The aim of this study was to design and validate, as a proof of concept (PoC), an AI module to detect information useful for registries in the classification and labeling of hip and knee implants, for inclusion within the RIPI framework.

2. Materials and Methods

The idea of conceiving an AI module emerged within the project "Creation of the prototype of the National Registry of Implantable Prostheses (RIPI) data collection platform (π-RIPI)" aimed at improving the registry IT infrastructure. The project was coordinated by ISS in collaboration with Confindustria Dispositivi Medici [27] and BVTech [28] and developed within the Rome Technopole framework, funded by the European Union [29].

The study to develop the AI module as PoC was organized in two phases: the first involved the design of the operational flow of the entire AI module, and the second involved the AI model development and validation in a laboratory setting.

2.1. The π-RIPI Context

The π-RIPI project aimed at implementing a unified, modular, and interoperable digital platform to centralize and harmonize the management of multiple implantable MD registries [30]. It was designed as a cloud-native, microservice-based digital architecture integrating modular interoperable components in compliance with privacy and security standards defined by national health data governance regulations.

2.2. Study Design

This study was designed as a PoC with a test application in a real-world framework. Its primary objective was to retrospectively evaluate the performance of an AI classification system by analyzing joint implant attributes, registry taxonomies, and information available in technical documents, without any prospective intervention. The study followed the STROBE guidelines for observational research, where applicable [31].

2.3. Design of the AI Module-PoC Operational Flow

To define the operational flow of the AI module-PoC, the models proposed by Ramachandran et al. [32] and Ren and Caiani [26] were considered suitable for application in this explored domain. Indeed, the approach by Ren and Caiani [26], which developed a web scraping and NLP framework for collecting MD adverse events, was selected as a reference for the AI module-PoC infrastructure in this study. Furthermore, the architectural design of the AI model, implemented in the AI module-PoC flow, was methodologically inspired by the hybrid Named Entity Recognition (hNER) approach proposed by Ramachandran et al. [32]. This approach used entity extraction from biomedical literature through the integration of domain-specific dictionaries and machine-learning models.

Both approaches were re-engineered to accommodate the architectural constraints and requirements of the π-RIPI infrastructure [30], with a focus on hip and knee implantable devices. The AI module-PoC was designed according to criteria of efficiency, traceability, and data quality, ensuring seamless integration into existing workflows while maintaining high standards of performance and reliability.

2.4. AI Module-PoC Development and Validation

The development and validation of the AI module-PoC were organized in the following four steps: pre-trained AI model selection, setup of the Custom-built dataset and model fine-tuning, AI model validation, and AI model external testing. In this pilot phase, web scraping capabilities were not included and technical documents that had already been collected by RIAP in .pdf format were used to train and test the AI models.

All phases were implemented using Python (Version 3.12.2, Python Software Foundation, Wilmington, DE, USA).

2.4.1. Pre-Trained AI Model Selection

The features of two pre-trained models integrated into a spaCy pipeline, i.e., Convolutional Neural Network (CNN) and Bidirectional Encoder Representations from Transformers optimized for biomedical text (BioBERT), were compared on the basis of the following specific project requirements: open-source software, limited computational effort, minimal reliance on local storage, no use of external cloud services, and acceptable accuracy levels. In the case of the CNN model, the pre-trained model en_core_web_sm (hereinafter referred to as "CNN" or "CNN model") was selected as it was deemed the most appropriate, having already been pre-trained using written web text and achieved an accuracy of 84% for the NER task across all relevant metrics in this phase (recall, precision, and F-score) [33]. As for the BioBERT model, the pre-trained model dmis-lab/biobert-base-cased-v1.1 (hereinafter

referred to as "BioBERT" or "BioBERT model") [34], was selected as it had already been pre-trained on large-scale biomedical text corpora and demonstrated high performance on biomedical NER benchmarks, with a reported accuracy of 98.49% according to the official model documentation [35].

2.4.2. Setup of the Custom-Built Dataset and Model Fine-Tuning

Figure 1 shows the setting up of the Custom-built dataset used to train and validate both models.

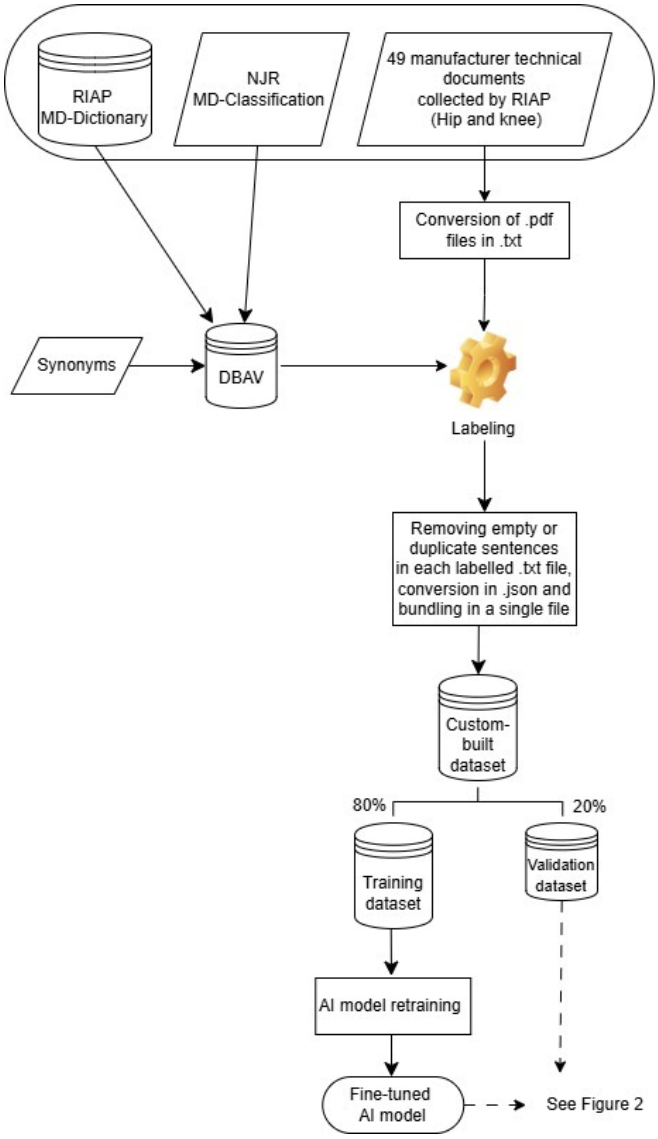


Figure 1. AI models development: flowchart for the setting up of the Custom-built dataset used to train and validate both AI-selected models. DBAV: Data Base of Attributes and Values.

The Custom-built dataset was set up using the following data sources as a reference: the RIAP MD-Dictionary (for structural and content information), the MD-Classification adopted by the National Joint Registry (NJR) for their Component-Database, and 49 technical documents concerning hip or knee prostheses that had already been collected by RIAP for activities related to the RIAP MD-Dictionary update. The full list of these technical documents (28 for hip components from 16 manufacturers, 21 for knee components from 13 manufacturers) is available as Supplementary Materials (Table S1). All the selected technical documents are written in English.

The following information was obtained from the RIAP MD-Dictionary and the NJR MD-Classification:

- Identification attributes. These enable the unambiguous identification of implantable devices. For the 3 attributes "Manufacturer", "Product code", and "European Medical Device Nomenclature code" (EMDN), all the values available in the RIAP MD-Dictionary were considered.
- Characterization attributes. These are essential for the technical characterization of each device. From the NJR MD-Classification, the attributes "Joint type", "Class Component", and "Class attribute" were considered with their values, except when they were binary (i.e., yes/no), related to a pre-defined numerical list (i.e., dimensions such as length, width, height, diameter, or size), or described in a generic or undefined way (i.e., the expressions other/other description at the end of a specific detailed list, N/A, None).

A specifically structured table, the MD Attributes/Values DataBase (DBAV), was created to implement both AI models. It consisted of 40 columns covering the 3 identification attributes (Manufacturer, Product code, and EMDN), the "Joint type" and the "Class Component" for both hip and knee, as well as the 35 selected "Class attributes". The first 3 columns were populated with all the values from the RIAP MD-Dictionary, while the remaining 37 were populated with the class values related to each attribute. The list of values in each column was supplemented with appropriate synonyms selected based on the knowledge of the research group.

The list of all the attributes considered in the DBAV is available as Supplementary Materials (Table S2). For reasons of confidentiality, the values associated with the class attributes are not presented.

All the technical documents collected by RIAP were converted from .pdf to .txt format using the Python library *pdfplumber*. Based on the experience of the research group, this library was considered the most effective for handling both text and tables with minimal information loss based on the following three main criteria: (1) sentence integrity, ensuring accurate extraction without word loss; (2) table recognition, with correct extraction of structured data while preserving cell-level positioning; and (3) line-level segmentation, ensuring one sentence per line.

Since some of the available information in the technical documents was in image format, the *pytesseract* Python library was incorporated into the processing pipeline to perform Optical Character Recognition (OCR). This integration enabled the automated extraction of textual content from rasterized sources, thereby ensuring their inclusion in downstream computational analyses.

To improve the line-level segmentation, an ad hoc function was developed to recognize the sentences by identifying specific punctuation marks (colon ":", semicolon ";", and full stop "."), followed by a blank space. Then, the developed algorithm scanned each sentence-containing line in each resulting .txt file. For each sentence, it compared each word with all the values listed in the DBAV. When a correspondence was found, the highlighted word in the sentence was labeled with the corresponding attribute. To improve the quality of the automatic labeling, the output of this phase was manually checked and words of interest available in the technical documents, but not already labeled by the algorithm, were added to the list of synonyms before starting a new iteration. Empty and duplicate sentences in each labeled .txt file were removed. Each .txt file, corresponding to a labeled technical document, was then converted into .json format. All the 49 .json files were then bundled in a single file resulting in the Custom-built dataset. This dataset was finally randomly partitioned into two sub-datasets: 80% of records were used to fine-tune the two pre-trained AI models, whilst the remaining 20% was allocated for models' validation. Partitioning

was performed using a sentence-level random split, with each sentence assigned to a single subset only. The CNN model was fine-tuned for 100 epochs with a dropout rate of 0.2 and learning rate of 0.001, whereas the BioBERT model was fine-tuned for 5 epochs with a dropout rate of 0.1 and learning rate of 0.00003. The different number of training epochs is related to the architecture of the two models. BioBERT is a transformer model already pre-trained on large biomedical text corpora and was therefore adapted to the specific NER task while the spaCy CNN model required a larger number of training iterations to reach stable performance under the adopted configuration.

2.4.3. AI Model Validation

The validation process adopted for both models (CNN and BioBERT) was structured in two phases (Figure 2).

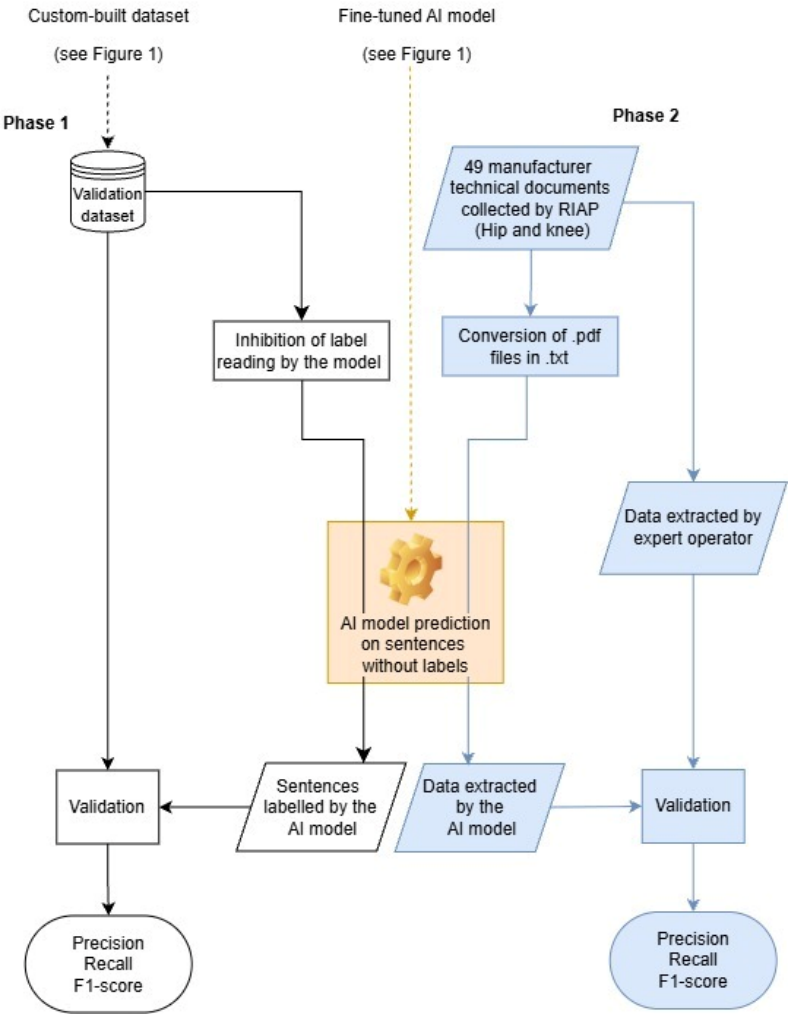


Figure 2. Flowchart of the AI model validation for labeling capacity of the 49 technical documents. Phase 1 (white): towards the automatic labeling; and Phase 2 (light blue): towards the labeling of an expert operator.

Phase 1 (Figure 2, white section of the flowchart) consisted of a comparison between the labeling already available in the validation dataset and labeling produced by AI models on the same sentences. Therefore, prior to providing these sentences to the AI models as input for prediction, their label reading function was inhibited. Confusion matrices were used to present the results of this comparison.

The following standard metrics were used to assess the ability of the models to correctly recognize and classify individual entities: precision (or positive predictive value), recall (or sensitivity), and F1-score. The equations are available as Supplementary Materials (Equations file S1). To account for class imbalance, Macro-F1 and Weighted-F1 scores were additionally computed.

In Phase 2 (Figure 2, light blue section of the flowchart), an expert operator (EO) was asked to manually label all the 49 technical documents by identifying the following 5 attributes: "Manufacturer", "Joint type", "Class Component", "Product code", and "Material". The selection grid used by the expert operator is available as Supplementary Materials (Excel S1).

The reference information was based on manual extraction performed by a single EO and was subsequently reviewed by two additional experts with experience in medical device classification and registry activities to ensure the quality and coherence of extracted information. The experts verified the completeness, accuracy, and consistency of the information extracted by the EO, providing an additional quality control step. EO labeling followed deterministic rules and was not left to personal interpretation; therefore, discordance was not expected in repeated independent assessment by the same or different operators. For each attribute, the values extracted manually by the expert operator were used as the reference and then compared with those automatically extracted by each AI model. For each technical document, the EO was asked to determine only the presence of an attribute, without counting the number of occurrences. In such a phase, the model was tested to recall the right attribute in the right context. The validation was performed by document (not by sentence) and "hip" or "knee" had to be identified as class component not only if mentioned, but if the document targeted that joint; for example, in a technical document related to a hip prosthesis, if the word "knee" appeared as related to the context of surgical technique and was identified by the model as class component, then it was considered as a misclassification.

As "Material" was not a specific attribute, the AI model was requested to extract the values associated with the following attributes: "Articulating Surface Modification", "Bone-Facing Coating", "Bone-Facing Composition", "Composition of Articulating Substrate", and "Substrate/Material". The following occurrences were considered for this assessment (see Venn diagram in Figure 3): the number of elements extracted by the expert operator (A) and, for each AI model, the number of elements extracted (B), the number of elements correctly extracted by the model ($A \cap B$), the number of elements extracted by the expert operator but not by the AI model ($A \setminus B$), and the number of elements not extracted by the expert operator but extracted by the AI model ($B \setminus A$).

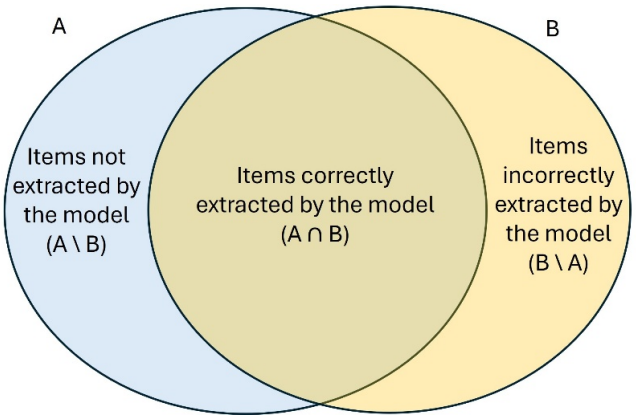


Figure 3. Venn diagram of the items extracted by the expert operator (A, light blue) and by each model (B, yellow).

All non-matching information was examined in detail to highlight aspects of the model that could be further improved.

For each technical document, the time required by the expert operator to manually extract the information and the time taken by the AI model to perform the same task were recorded. Finally, the average and standard deviation values of the data analyzed across all 49 technical documents were computed.

2.4.4. AI Model External Testing

The external testing phase was conducted using a set of 25 technical documents (hip and knee prostheses), all of which were unknown to both AI models (Figure 4). The manufacturer overlap between the training/validation and external-test datasets was 81.8% (9 out of 11 manufacturers). The implant devices represented in the external-test dataset were entirely different from those included in the training and validation dataset, and no device-specific technical document was shared between the two datasets.

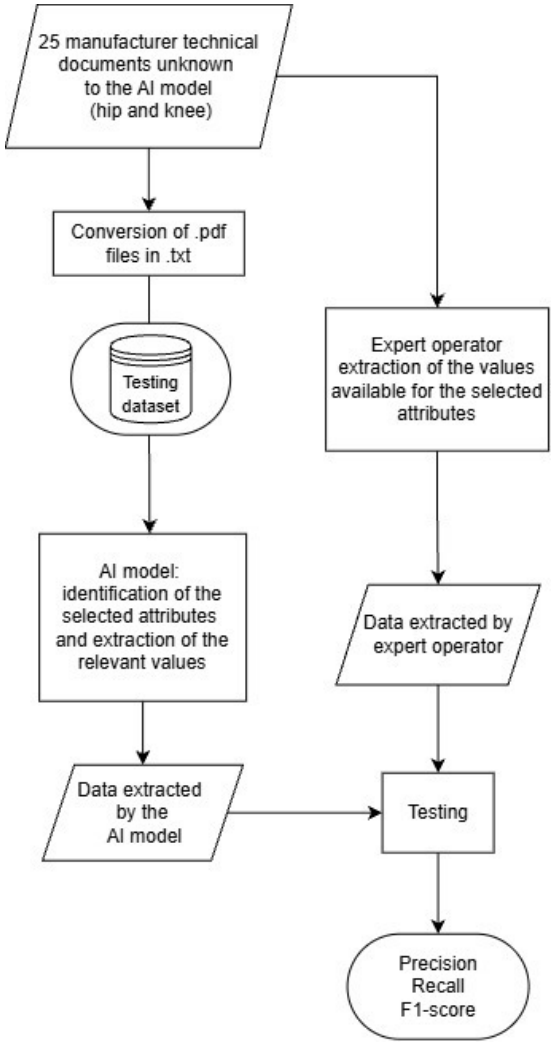


Figure 4. Flowchart of the AI model external testing: identification of the selected attributes ("Manufacturer", "Joint type", "Class Component", "Product code", and "Material") and of the relevant associated values.

All the technical documents in .pdf format were converted into .txt files to ensure standardized processing and subsequently compiled into a dedicated external testing dataset. The two AI models processed the dataset autonomously, performing automatic data recognition and extraction according to the predefined information fields. In parallel,

an expert operator manually reviewed the same set of technical documents and extracted the required information following the same criteria used for the AI models. The outputs produced by the AI models were compared with those extracted by the expert operator for the same 5 attributes: "Manufacturer", "Joint type", "Class Component", "Product code", and "Material". The performance of each model was quantified by computing precision, recall, and F1-score.

2.5. Ethics

The study used only data concerning MDs (product code, manufacturer name, and technical attributes). No patient data was used in this study; therefore, formal ethical approval was not necessary.

3. Results

3.1. AI Module-PoC in the π-RIPI Context

Figure 5 shows how the AI module-PoC will be integrated as an additional component of the π-RIPI platform.

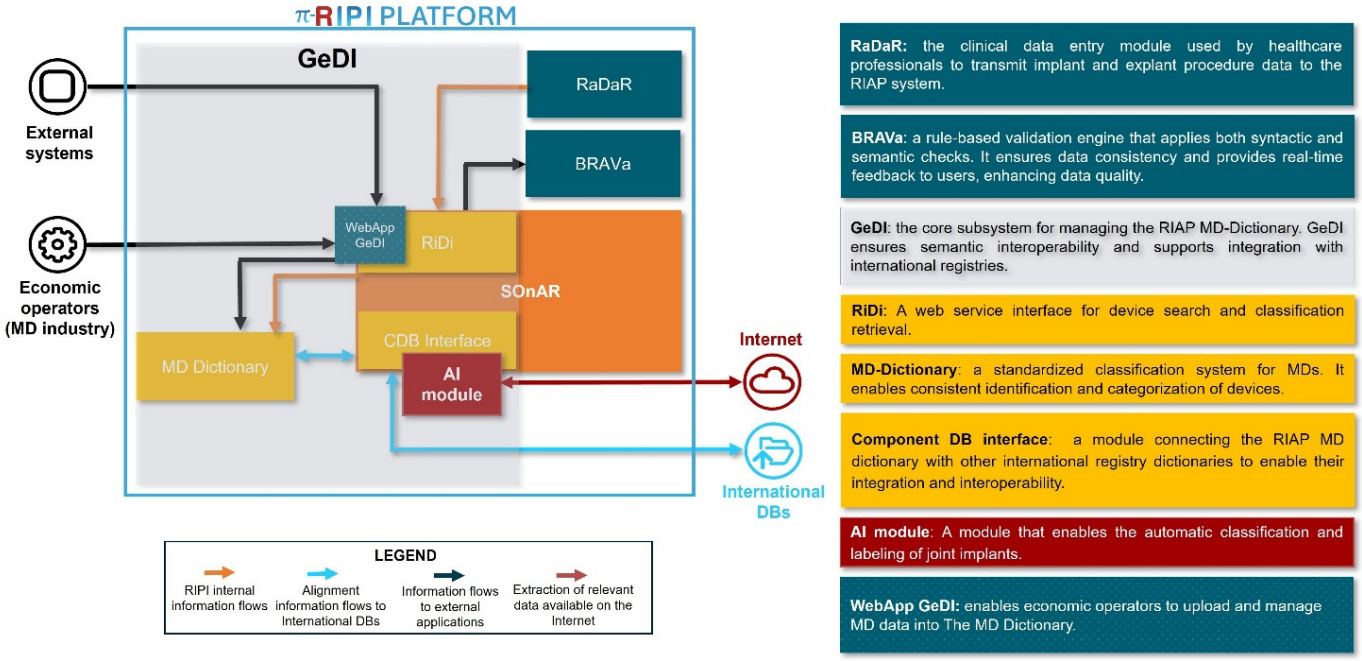


Figure 5. Inclusion of AI module-PoC in π-RIPI and description of the platform components.

3.2. AI Module-PoC Operational Flow

Figure 6 shows the operational flow of the future AI module-PoC once it is deployed. In the present study, only stages 2 and 3 were conducted. In stage 1, only the .pdf technical documents that had previously been collected by RIAP were considered. As soon as the production phase of the AI module-PoC is launched, stage 1 will encompass all web scraping activity. The operational flow of the entire AI module-PoC is illustrated in detail as Supplementary Materials (Figure S1) and encompasses seven main stages: Web scraping, Data processing, Task submission, Ranking & Results validation, Results Aggregation, Admin Validation, and Upload on to the MD-Dictionary.

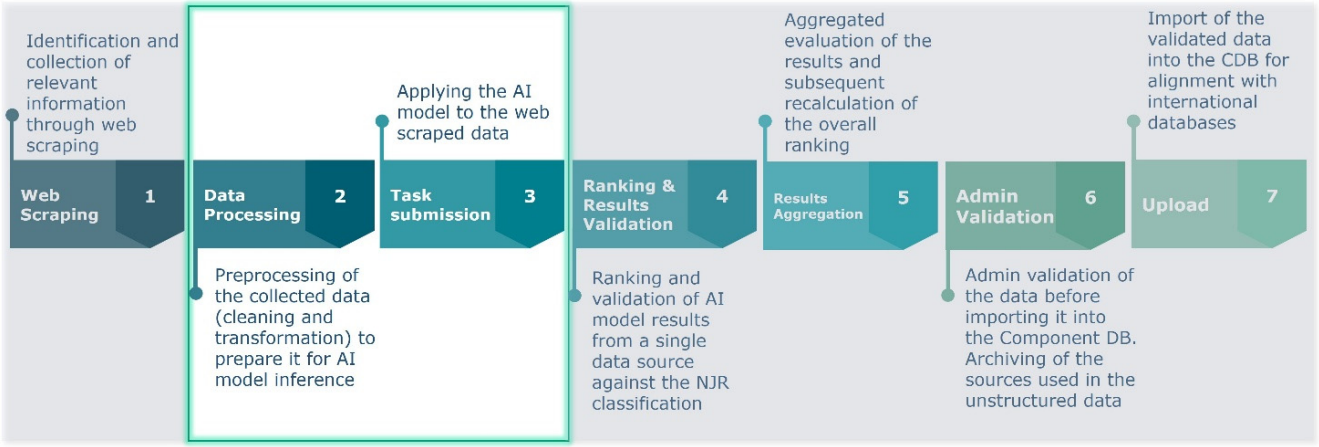


Figure 6. Operational flow of the future AI module-PoC. Stages 2 and 3 developed in this study are highlighted.

3.3. AI Module-PoC Development and Validation

3.3.1. Setup of the Custom-Built Dataset and Model Training

The following attributes of the NJR MD-Classification (taxonomy) were excluded to implement the AI module-PoC: 8 attributes whose values were binary (2 values), 4 attributes whose values were related to a pre-defined numerical list (50 values), and 11 attributes whose values were described in a generic or undefined way (5 values) (Table S2. List of all selected attributes from RIAP MD-Dictionary and NJR MD-Classification). The following attributes were included in the DBAV: 3 identification attributes, described by 117,757 distinct values, from the RIAP MD-Dictionary; and 37 characterization attributes, described by 299 distinct values, from the NJR MD-Classification. Since EMDN is derived from the Italian CND, this last one was included in the DBAV as a synonym for the EMDN attribute. Moreover, synonyms of 122 values included in the NJR MD-Classification were added to the DBAV. Finally, the DBAV included 40 attributes and 118,179 values.

After removing empty or duplicate sentences, the labeling of the 49 technical documents resulted in 13,613 sentences (6770 for hip and 6843 for knee devices, length ranging from 1 to 1433 characters) that were then uploaded in the Custom-built dataset. Table 1 reports the distribution of the number of labeled sentences organized by class of length, where the length is the number of characters, together with the average and standard deviation.

Table 1. Number of labeled sentences, and average and standard deviation of length, by class of length.

Class	n	Average Length	Standard Deviation
1–80	10,091	42.25	19.91
81–160	2706	110.49	21.24
161–240	593	189.19	21.45
241–320	86	271.33	20.85
321–400	50	353.36	23.00
>400	87	554.05	158.21

Most sentences are concentrated in the 1–80 class, with 10,091 occurrences. The number of sentences decreases progressively to 50 occurrences in the class 321–400 and increases up to 87 occurrences in the 401–1433 class, which has the highest standard deviation. In this last class, the labeling algorithm allocated the text presented in the technical documents in the form of tables.

Table 2 shows the list of the 34 attributes identified during the labeling phase of the 49 technical documents and included in the Custom-built dataset.

Table 2. Custom-built dataset: attributes identified in the 49 technical documents analyzed during the automatic labeling phase.

Attribute	
Identification	
EMDN/CND	Product code
Manufacturer	
Characterization	
Anchoring/Fixation	Joint type
Articulating Shape Design	Location
Articulating Surface Modification	Modification
Bone-Facing Coating	Modularity
Bone-Facing Composition	Neck Length
Bone-Facing Texture	Polyethylene Modification
Bone Interface Design	Primary Fixation
Class Component	Relationship to Patella
Composition of Articulating Substrate	Secondary Fixation
Constraint	Shape
Cross-Section	Side
Degree of Congruency	Substrate/Material
Degree of Stability	Taper
Design	Taper Texture
Fixation Surface Texture	Type
Insert Interface Design	

3.3.2. AI Model Validation

Figure 7 shows the confusion matrices of the 20 most frequent attributes that were correctly classified by each model during the labeling in Phase 1 validation. In such confusion matrices, rows represent the true labels and columns represent the predicted ones.

Out of the 34 attributes included in the Custom-built dataset, 19 were correctly identified more than 20 times by the CNN (Figure 7a) model and 20 by the BioBERT model (Figure 7b). Complete confusion matrices are provided as Supplementary Materials (Excel S2). In total, the CNN correctly identified 26 attributes, while BioBERT correctly identified 30.

The comparison between the two confusion matrices highlights an improvement in classification performance when using the BioBERT model compared to the CNN-based model. In terms of correctly classified instances (diagonal values), both models performed well for the two most represented attributes "Product code" and "Class Component" (CNN achieved 1135 and 1056 correct predictions for these attributes, respectively, and BioBERT achieved 936 and 1404, respectively). Table 3 shows the results of the three metrics, precision, recall, and F1-Score, obtained for both AI models for every attribute.

Excel S2 shows the complete confusion matrices, not restricted to the top 20 items, for the models CNN and BioBERT, respectively. The last row accounts for spurious prediction, contributing to false positives, whereas the last column accounts for target items missed by the model, contributing to false negatives.

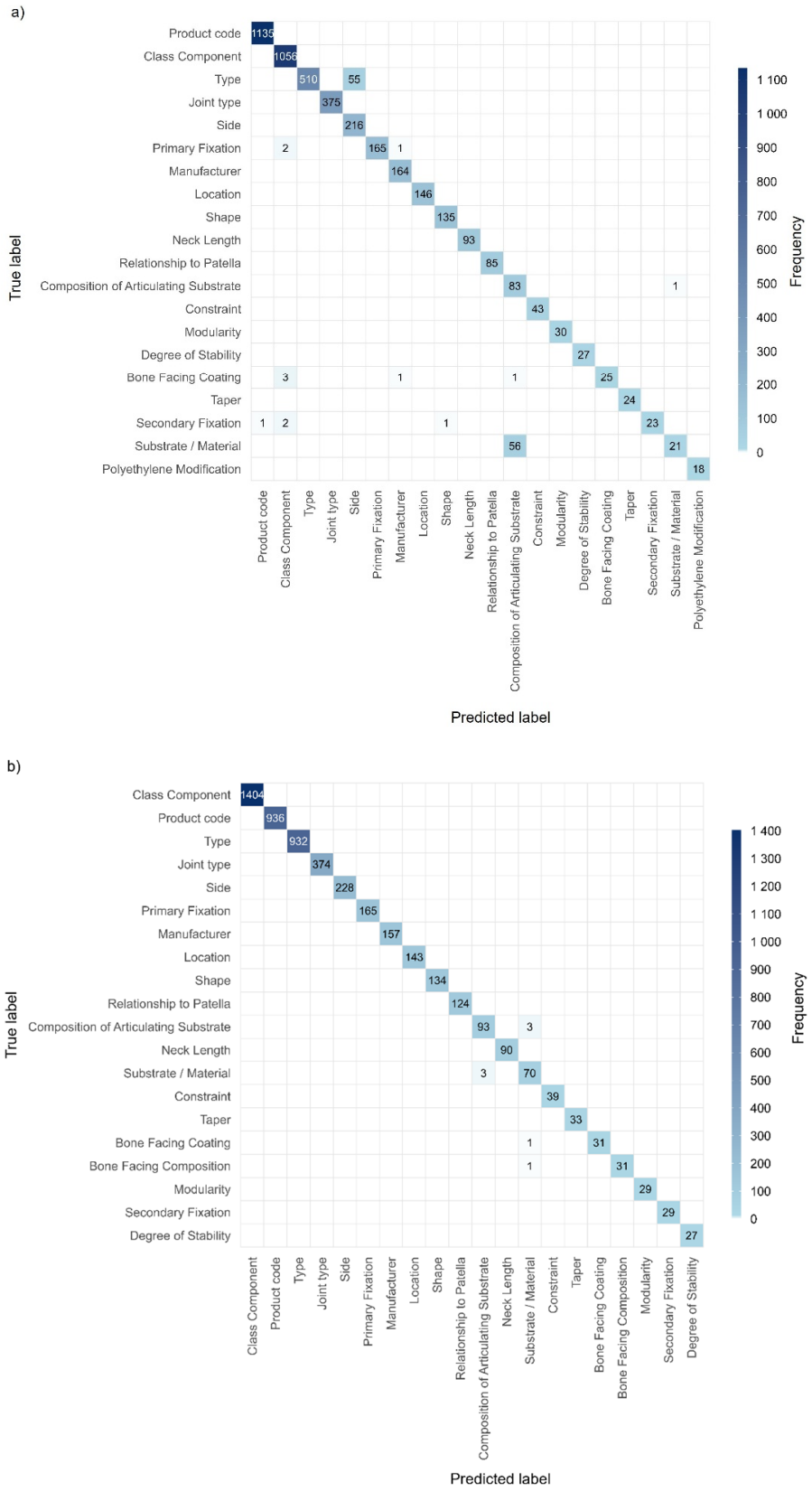


Figure 7. Confusion matrices of entity labels for the 20 most frequent attributes used to evaluate the performance of the CNN model (**a**) and the BioBERT model (**b**). The y-axis represents the attributes as available in the technical documents (true), while the x-axis corresponds to the attributes predicted by each AI model. Each cell in the confusion matrix indicates the number of times an entity with a given true label was classified as a specific predicted label. The heatmap bar reflects the frequency of occurrences: the darker the color, the higher the number of instances represented in that cell.

Table 3. Performance metrics of both models for each of the 34 attributes included in the Custom-built dataset.

Attribute	Precision		Recall		F1-Score	
	CNN	BioBERT	CNN	BioBERT	CNN	BioBERT
Anchoring/Fixation	NP	100	0	100	NP	100
Articulating Shape Design	0	100	0	100	NP	100
Articulating Surface Modification	NP	100	0	100	NP	100
Bone-Facing Coating	34.25	96.88	71.43	88.57	46.30	92.54
Bone-Facing Composition	100	100	15.15	93.94	26.32	96.88
Bone-Facing Texture	0	NP	0	0	NP	NP
Bone Interface Design	83.33	100	55.56	88.89	66.67	94.12
Class Component	64.55	97.84	72.78	96.76	68.42	97.30
Comp. of Articulating Substrate	42.35	90.29	79.05	88.57	55.15	89.42
Constraint	81.13	90.70	100	90.70	89.58	90.70
Cross-Section	100	33.33	60.00	20.00	75.00	25.00
Degree of Congruency	77.78	77.78	100	100	87.50	87.50
Degree of Stability	61.36	100	100	100	76.06	100
Design	0	NP	0	0	NP	NP
EMDN	NP	NP	0	0	NP	NP
Fixation Surface Texture	37.50	75.00	33.33	66.67	35.29	70.59
Insert Interface Design	NP	0	0	0	NP	NP
Joint type	84.08	99.73	98.68	98.42	90.80	99.07
Location	67.91	89.94	94.81	92.86	79.13	91.37
Manufacturer	68.91	96.91	91.11	87.22	78.47	91.81
Modification	100	83.33	80.00	100	88.89	90.91
Modularity	71.43	90.62	96.77	93.55	82.19	92.06
Neck Length	43.46	65.22	40.26	38.96	41.80	48.78
Polyethylene Modification	60.00	94.12	100	88.89	75.00	91.43
Primary Fixation	79.71	97.63	97.06	97.06	87.53	97.35
Product code	76.43	91.67	89.30	73.64	82.37	81.68
Relationship to Patella	75.22	97.64	67.46	98.41	71.13	98.02
Secondary Fixation	69.70	90.62	71.88	90.62	70.77	90.62
Shape	65.22	86.45	95.74	95.04	77.59	90.54
Side	52.94	93.83	91.53	96.61	67.08	95.20
Substrate/Material	67.74	85.37	26.92	89.74	38.53	87.50
Taper	52.17	100	70.59	97.06	60.00	98.51
Taper Texture	NP	100	0	25.00	NP	40.00
Type	78.46	98.21	53.07	96.98	63.31	97.59

Legend: NP not possible because it is an indeterminate form.

Both models were evaluated on the same test set, which included 5810 annotated target items. However, the total number of evaluated pairs in the confusion matrices exceeds such a number, as the matrices also include model predictions that do not correspond to any target item. The total number of evaluated pairs was 6090 for BioBERT, including 280 incorrect identifications of items not of interest, and 7655 for CNN, including 1845 incorrect identifications of items not of interest. This difference indicates that CNN generated a substantially higher number of spurious predictions than BioBERT. Therefore, the higher number of incorrect identifications for CNN indicates a greater tendency to generate non-relevant spans, compared to BioBERT (Excel S2. Complete confusion matrices for both validated models).

Table 3 shows the values obtained for the standard metrics used to assess the model's predictive ability for each of the 34 attributes included in the Custom-built dataset.

In general, Recall, which measures the actual model's ability to identify attributes, is comparable between the two models. Nonetheless, when considering the three most frequent attributes, BioBERT outperforms the CNN in identifying "Type" and "Class component", with Recall values of 96.98% and 96.76% against values equal to 53.07% and

72.78%, respectively. Conversely, the CNN outperforms BioBERT in identifying "Product code", with a Recall value equal to 89.30% against 73.64%.

Additional performance metrics accounting for the class imbalance showed Macro-F1 and Weighted-F1 values of 68.50% and 71.68% for the CNN model, and 87.55% and 90.97% for the BioBERT model, respectively.

Table 4 reports the results of standard evaluation metrics obtained on the validation dataset for the two AI models tested.

Table 4. Standard evaluation metrics obtained by the two AI models in the Phase 1 validation.

Metric	AI Model	
	CNN Model	**BioBERT Model**
Precision	67.18%	94.62%
Recall	75.82%	88.45%
F1-score	71.24%	91.43%

The BioBERT model achieved higher values across all evaluated performance metrics compared to the CNN-based model, including Precision, Recall, and F1-score.

Table 5 presents the aggregated results of the Phase 2 validation, i.e., the extractions of the 49 technical documents performed by the expert operator and by the two AI models. The detailed results of the comparison for each technical document are provided as Supplementary Materials (Excel S3).

In the validation test, BioBERT shows higher values of recall than the CNN for the following three attributes: "Product code" (55.8% vs. 46.1%), "Joint type" (95.9% vs. 83.7%), and "Class Component" (54.2% vs. 24.9%). For "Material", the performance of the two models is similar (46.5% for BioBERT, 47.2% for CNN), while, for "Manufacturer", the recall of CNN is higher than that of BioBERT (82.5% vs. 73.7%).

BioBERT shows lower FN values than the CNN for the "Product code" (3377 vs. 4115) and "Class Component" (81 vs. 133). In terms of precision rates, BioBERT shows higher values for the "Class Component" (53.9% vs. 28.9%), "Product code" (84.9% vs. 78.6%), and "Material" (88.0% vs. 77.0%).

In addition, both AI models showed a substantial reduction in processing time compared to the expert operator. The expert operator needed a total of 79,980 s extracting the 49 technical documents manually, while CNN completed the same task in 275 s, and BioBERT in 1600 s. This corresponds to an average time per document of 1632.24 s for the expert operator, 5.61 s for the CNN, and 32.66 s for BioBERT.

3.3.3. AI Model External Test

Table 6 presents the results of the AI model external testing.

Both models achieved a high level of recall in identifying the "Joint type" (92%). For the other attributes, the following values of recall are measured: "Manufacturer", 53.6% for BioBERT and 28.6% for the CNN; "Product code", 34.0% for BioBERT and 11.0% for the CNN; "Class Component", 54.2% for BioBERT and 50.0% for the CNN; and "Material", 42.9% for BioBERT and 19.6% for the CNN.

BioBERT shows lower FN values than the CNN for the "Product code" (2010 vs. 2711) and "Class Component" (33 vs. 36). In terms of precision rates, the CNN shows higher values than BioBERT for the "Manufacturer" (88.9% vs. 65.2%), "Product code" (92.3% vs. 73.7%), and "Class Component" (48.0% vs. 43.8%), the same value for "Joint type" (76.7%), and a lower value for "Material" (84.6% vs. 96.0%).

Table 5. Results of Phase 2 validation test: comparison between automatic (CNN and BioBERT models) and EO manual labeling of 49 technical documents for 5 selected attributes.

	Elements Extracted			TP		FN		FP		Recall		Precision		F1-Score	
	EO	**CNN**	**BioBERT**	**CNN**	**BioBERT**	**CNN**	**BioBERT**	**CNN**	**BioBERT**	**CNN**	**BioBERT**	**CNN**	**BioBERT**	**CNN**	**BioBERT**
TOTAL															
Manufacturer	57	62	65	47	42	10	15	15	23	82.5%	73.7%	75.8%	64.6%	79.0%	68.9%
Product code	7641	4484	5025	3526	4264	4115	3377	958	761	46.1%	55.8%	78.6%	84.9%	58.2%	67.3%
Joint type	49	46	54	41	47	8	2	5	7	83.7%	95.9%	89.1%	87.0%	86.3%	91.3%
Class Component	177	152	178	44	96	133	81	108	82	24.9%	54.2%	28.9%	53.9%	26.7%	54.1%
Material	142	87	75	67	66	75	76	20	9	47.2%	46.5%	77.0%	88.0%	58.5%	60.8%
Time (s)	79,980	275	1600												
MEAN															
Manufacturer	1.16	1.27	1.33	0.96	0.86	0.20	0.31	0.31	0.47						
Product code	155.94	91.51	102.55	71.96	87.02	83.98	68.92	19.55	15.53						
Joint type	1.00	0.94	1.10	0.84	0.96	0.16	0.04	0.10	0.14						
Class Component	3.61	3.10	3.63	0.90	1.96	2.71	1.65	2.20	1.67						
Material	2.90	1.78	1.53	1.37	1.35	1.53	1.55	0.41	0.18						
Time (s)	1632.24	5.61	32.66												
SD															
Manufacturer	0.43	1.04	1.14	0.54	0.61	0.46	0.47	0.82	0.89						
Product code	207.84	154.82	174.17	132.79	154.09	139.21	130.06	41.20	31.10						
Joint type	0.00	0.52	0.37	0.37	0.20	0.37	0.20	0.31	0.35						
Class Component	2.33	2.62	2.07	1.18	1.34	2.35	1.95	2.23	1.51						
Material	1.75	1.83	1.65	1.45	1.33	1.39	1.23	0.81	0.57						
Time (s)	1385.04	7.66	56.57												

Legend: EO = Expert Operator; TP = Elements correctly extracted by AI model (True Positives); FN = Elements extracted by EO but not by AI model (False Negative); FP = Elements not extracted by EO but by AI model (False Positive).

Table 6. Results of external test: comparison between automatic (CNN and BioBERT models) and EO manual labeling of 25 technical documents for 5 selected attributes.

	Elements Extracted			TP		FN		FP		Recall		Precision		F1-Score	
	EO	**CNN**	**BioBERT**	**CNN**	**BioBERT**	**CNN**	**BioBERT**	**CNN**	**BioBERT**	**CNN**	**BioBERT**	**CNN**	**BioBERT**	**CNN**	**BioBERT**
TOTAL															
Manufacturer	28	9	23	8	15	20	13	1	8	28.6%	53.6%	88.9%	65.2%	43.2%	58.8%
Product code	3046	363	1406	335	1036	2711	2010	28	370	11.0%	34.0%	92.3%	73.7%	19.7%	46.5%
Joint type	25	30	30	23	23	2	2	7	7	92.0%	92.0%	76.7%	76.7%	83.6%	83.6%

Table 6. *Cont.*

	Elements Extracted			TP		FN		FP		Recall		Precision		F1-Score	
	EO	**CNN**	**BioBERT**	**CNN**	**BioBERT**	**CNN**	**BioBERT**	**CNN**	**BioBERT**	**CNN**	**BioBERT**	**CNN**	**BioBERT**	**CNN**	**BioBERT**
Class Component	72	75	89	36	39	36	33	39	50	50.0%	54.2%	48.0%	43.8%	49.0%	48.4%
Material	56	13	25	11	24	45	32	2	1	19.6%	42.9%	84.6%	96.0%	31.9%	59.3%
Time (s)	36,120	23	544												
MEAN															
Manufacturer	1.12	0.36	0.92	0.32	0.60	0.80	0.52	0.04	0.32						
Product code	121.84	14.52	56.24	13.40	41.44	108.44	80.40	1.12	14.80						
Joint type	1.00	1.20	1.20	0.92	0.92	0.08	0.08	0.28	0.28						
Class Component	2.88	3.00	3.56	1.44	1.56	1.44	1.32	1.56	2.00						
Material	2.24	0.52	1.00	0.44	0.96	1.80	1.28	0.08	0.04						
Time (s)	1488.00	0.95	22.05												
SD															
Manufacturer	0.33	0.49	0.86	0.48	0.65	0.41	0.51	0.20	0.69						
Product code	167.87	28.20	114.70	27.14	82.93	154.33	133.57	2.67	33.79						
Joint type	0.00	0.58	0.58	0.28	0.28	0.28	0.28	0.46	0.46						
Class Component	2.20	1.22	1.39	0.77	0.82	1.78	1.63	1.08	1.35						
Material	1.36	0.87	1.00	0.77	1.02	1.08	1.02	0.28	0.20						
Time (s)	1347.33	0.84	21.16												

Legend: EO = Expert Operator; TP = Elements correctly extracted by AI model (True Positives); FN = Elements extracted by EO but not by AI model (False Negative); FP = Elements not extracted by EO but by AI model (False Positive).

A similar time pattern with respect to the validation test was observed. The expert operator took 36,120 s to process all documents, while the CNN model took 23 s, and the BioBERT model 544 s. The corresponding average processing times per document were approximately 0.95 s for the CNN model and 22.05 s for the BioBERT model, compared to 1488.00 s for the expert operator.

4. Discussion

This study provides preliminary evidence supporting the feasibility and initial effectiveness of integrating an AI-based classification system into a national MD registry dictionary. Specifically, it outlines the PoC implementation of the AI module and assesses the performance of its underlying model in analyzing real-world data related to hip and knee implants. Overall, the results support the potential of using NLP techniques, particularly NER, to extract structured information on MDs from unstructured sources, such as technical documents, to automatically populate the registry's MD-Dictionary. This approach could enhance interoperability, automate data collection, and enable advanced analytics for MD surveillance, in line with the best practices established by international registry models [13,15,16].

The results of this study also contribute to the growing body of research on AI applications in orthopedics. In recent years, AI has gained increasing attention in this field, not only for diagnostic purposes [22], but also for the automated extraction of structured data from unstructured clinical texts [36]. Karnuta et al. [37] developed a deep-learning model that accurately identified 18 hip implant designs from radiographs, underscoring the broader potential of AI in orthopedic device tracking. A recent review [19] identified 24 studies applying NLP to orthopedic areas such as diagnosis, adverse event detection, auditing, coding, and patient experience analysis. Aydın and Orhan [21] further confirm the rapid expansion of AI in orthopedics and its impact on patient care. However, none of these works focus on the automatic and fast extraction of MD data directly from manufacturer websites and product technical documents. This procedure has the potential to greatly enhance the role of registries in supporting post-market surveillance. To our knowledge, this is the first study to explore this application. However, although focused on a different task, Ren and Caiani [26] demonstrated the potential of NLP by classifying over 65,000 field safety notices from 16 EU countries, highlighting the challenge of managing heterogeneous data and the implications for the standardization and interoperability of registries. Their findings underscore the potential of automated approaches to substantially enhance the data standardization, traceability, and post-market surveillance of MDs within a highly fragmented regulatory framework. The AI module-PoC developed in this study follows a methodological framework similar to that used by Ren and Caiani [26].

The architectural integration of the AI module-PoC into the π-RIPI platform—currently limited to the RIAP registry—is a fundamental step towards implementing the entire RIPI automated ecosystem for MD surveillance. At this stage, web scraping has not yet been implemented; instead, already available technical documents were used to build the dataset. The premature implementation of web scraping, without confirming the data availability, would have hindered the ETL (Extract, Transform, and Load) pipeline design. The operational flow tested in the PoC was limited to stages 2 and 3, corresponding to only a subset of the future module. However, it already reflects the essential components of the full workflow: data ingestion, processing, task submission, and preliminary validation. This structured design will ensure future scalability once the web scraping capabilities and higher-order inference stages are activated. Although the PoC only used previously collected PDFs, the planned operational flow provides a robust foundation for the future deployment of the system in a fully automated way. Ongoing advancements are focusing

on developing a robust web-scraping algorithm capable of handling dynamic content, CAPTCHA, and complex site structures to enable comprehensive data acquisition.

One of the major challenges at the outset of the project was the lack of a comprehensive, unified dataset. Integrating heterogeneous sources such as the RIAP MD-Dictionary, the NJR MD-Classification, and manufacturer technical documents into a harmonized dataset enabled the model to generalize across different terms and formats. Moreover, several synonyms were added to the DBAV. This approach addressed terminological inconsistencies and incomplete labeling, contributing to a more accurate and harmonized device classification. This strategy may have contributed to achieving a promising performance in both the validation and external testing phases. The developed dataset includes identification and characterization variables: the former allows MD traceability, while the latter enables comparative survival and epidemiological analyses. For this study, labels such as "NaN" or "Other", ambiguous numerical values (e.g., diameter vs. length), and binary classes (yes/no) were excluded from automatic labeling in order to preserve data quality. These exclusions were methodological, not conceptual. Future extensions to this AI module-PoC will be required to manage structured numeric attributes, hierarchical fields, and manufacturer-specific terminology. Since the AI models were trained to perform NER at the sentence level, the Custom-built dataset was partitioned into subsets using individual sentences as the unit of analysis. As a result, while each sentence was assigned to a single subset, sentences originating from the same technical document could be distributed across the different subsets. To further assess the model performance on previously unseen data, an independent external test based on 25 technical documents not included in the dataset partitioning process was also performed. Future studies may explore document-level partitioning strategies.

The Custom-built dataset, which included the information from the RIAP Dictionary and the NJR MD-Classification, in addition to the 49 technical documents from 29 manufacturers, provided a solid starting point but did not fully capture the variability of synonyms and terminology variation used in manufacturers' technical documents. This dataset directly influenced the downstream performance despite representing only a preliminary foundation. This was particularly evident in the external testing, especially for attributes such as "Class Component" and "Material", where the linguistic variation and context dependency are more pronounced. Most of the misclassifications were attributed to missing synonyms, lexical variants (such as hyphenated manufacturer names), or contextual indications that the AI models were not yet capable of interpreting. This reinforces the need for larger, more diverse training sets and expanded synonym sets, including manufacturer-specific nomenclature and registry taxonomies.

The distribution of sentence lengths in the analyzed technical documents aligns with typical patterns observed in natural language, with most sentences shorter than 80 characters. This supports the suitability of sentence-level NER approaches, although longer or fragmented sentences, especially in multicolumn layouts, remain a challenge. Short sentences may lack sufficient contextual information for accurate interpretation, whereas excessively long sentences increase syntactic complexity and processing difficulty. In this context, using the *pdfplumber* library proved effective for extracting tables, though it has limitations in handling complex PDF structures, which may result in misclassification. Importantly, the integration of OCR improved the system's ability to handle rasterized technical documents, addressing a major limitation encountered in earlier versions. Nevertheless, multicolumn layouts, merged pages, and embedded tables still challenged segmentation, occasionally producing overlong sentences (>400 characters) that impaired both sentence-level parsing and accurate label assignment. Ongoing efforts focus on expanding the dataset to obtain a

solid base comparable with other similar studies [26,38,39], refining entity definitions and improving label harmonization across sources.

The frequency distributions reported for hip and knee attributes reveal the intrinsic imbalance within the available documentation. Attributes such as "Product code" and component-specific descriptors appear abundantly, whereas others, especially regulatory or taxonomy-related items such as EMDN, are rarely or never mentioned. The near-complete absence of EMDN in commercial technical documents highlights a structural limitation of registry-driven attribute extraction: manufacturers do not consistently reference standardized nomenclatures in their publicly available documents, despite recommendations at the European level [40].

Two different NLP approaches were explored in this study to evaluate the impact of the model architecture on the entity recognition performance within the same experimental framework: a CNN-based model and a BioBERT-based model. From a technical standpoint, the comparison between the CNN-based model and the transformer-based BioBERT model confirms the expected trade-off between computational efficiency and predictive performance. As emerged from the scientific literature, BioBERT, in general, offers superior contextual understanding and accuracy, whereas spaCy demonstrates significantly faster inference times (scispaCy processed an average sentence in approximately 90 ms, whereas BioBERT required around 278 ms per sentence) with only a marginal decrease in performance, as indicated by a slightly lower but still competitive F1-score [41]. Another study [42] comparing scispaCy and BioBERT for biomedical NER confirmed that scispaCy is faster and more lightweight, making it suitable for real-time or resource-constrained applications. In this study, the initial selection of the CNN model was motivated by its lower computational requirements and faster execution, but the direct experimental evaluation demonstrated that the BioBERT model consistently outperformed the CNN model across most attributes. In the validation test, BioBERT achieved higher recall values than the CNN model for key attributes such as the "Product code", "Joint type", and "Class Component", while maintaining a comparable performance for "Material". Similar trends were observed in the external test, where BioBERT showed clear improvements for "Product code", "Manufacturer", and "Material".

The CNN-based AI model achieved an F1-score of 71.24%, with a recall of 75.82% and slightly lower precision (67.18%), reflecting a balanced trade-off between precision and recall. In comparison, the BioBERT model demonstrated a higher overall performance, achieving an F1-score of 91.43%, with both precision and recall exceeding 88%, confirming its superior capability in capturing contextual information and reducing misclassification.

Although the performance of the CNN model is lower than that reported [38] (F1-score of 96.3% with an AI model designed to automatically extract clinical information from orthopedic operative reports), the results obtained with BioBERT are comparable. The observed differences can be attributed to several factors: (1) the dataset used in this study consists of heterogeneous PDF documents with complex layouts and highly variable terminology, whereas the comparator model was trained on clinical operative reports, which are typically more structured and lexically homogeneous; and (2) the models under study are required to manage a substantially larger number of classification categories, increasing the likelihood of ambiguity and misclassification, in contrast to the nine predefined clinical labels used in the other study.

Together with standard evaluation metrics, the confusion matrices provided additional insights into the model behavior for both the CNN-based and BioBERT models. The BioBERT confusion matrix appeared more concentrated along the main diagonal, indicating a higher degree of agreement between the true labels and predicted labels, while the CNN confusion matrix shows a broader dispersion of counts outside the main diagonal,

suggesting a greater fragmentation of predictions. This pattern indicated that the CNN was more prone to misclassifying entities across attribute classes, while also producing a larger number of false-positive predictions. Classes with a limited number of annotated examples showed a lower stability and were more frequently involved in misclassification patterns, particularly for the CNN. In contrast, BioBERT showed a more stable distribution of predictions across classes, with fewer wrongly identified items and a stronger concentration of values along the main diagonal. Indeed, the BioBERT confusion matrix indicates a more stable performance in handling the terminological variability present in the texts, although it shows errors and omissions. This result is consistent with the higher overall performance observed for the BioBERT model and with the lower number of spurious predictions compared with the CNN. When considering the classification of single attributes, those with a high performance, such as "Product code" and "Joint type", showed strong diagonal distributions in both matrices, reflecting a high classification consistency supported by lexical regularity. In contrast, misclassifications in both models tended to concentrate around attributes belonging to overlapping semantic domains, such as "Bone-Facing coating", "Composition of Articulating Substrate", and "Substrate/Material", or around attributes whose identification is highly context dependent, such as the distinction between augment components and surgical instruments. In addition, some categories were represented by a limited number of training examples, which contributed to the increasing variability in their classification. Across both models, a substantial proportion of errors appeared to be associated with incomplete synonym coverage rather than intrinsic architectural limitations. Performance patterns across the different classes generally reached more stable results for BioBERT than for the CNN model. For several high-frequency attributes characterized by relatively stable terminology, both models achieved a high performance; however, BioBERT more frequently maintained a better balance between precision and recall. The difference between the two models becomes more evident for less represented or semantically more variable classes. In these categories, the CNN model showed more pronounced fluctuations between precision and recall, including cases with null or non-computable values, reported as 0 or NP. This indicates a lower stability in recognizing rare or poorly represented classes, especially when the training set does not include enough examples to capture the different ways in which these entities are expressed in the documents. Although BioBERT also showed some limitations, it maintained a more stable predictive performance across a larger number of classes.

During the manual validation phase, additional synonyms were iteratively incorporated into the dataset, confirming the dynamic nature of terminology harmonization across manufacturers and the progressive refinement of the system. The manual processing of 49 technical documents (for the validation test) required more than 22 h, whereas the CNN-based model completed the task in about 4 min and the BioBERT model in about 26 min. This corresponds to a reduction in processing time of approximately 99.7% for the CNN model and 98.0% for the BioBERT model and quantifies the module's potential impact on registry scalability. Similar trends were observed in the external test, where the processing time decreased from about 10 h to 23 s with CNN and to about 9 min with BioBERT. These results demonstrate the system's ability to substantially accelerate large-scale document processing, addressing one of the most time-consuming activities of registry data management.

In the validation test, the highest recall values were observed for "Joint type" for both models, reaching 95.9% with BioBERT and 83.7% with the CNN model. In contrast, "Product code" showed moderate recall for both approaches, with BioBERT outperforming the CNN model (55.8% vs. 46.1%), reflecting the difficulty in extracting highly variable alphanumeric identifiers. Lower performance values were observed for the "Class Com-

ponent" and "Material", although BioBERT showed a clear improvement for the "Class Component" (54.2% vs. 24.9%). The results for "Material" remained comparable between the two models (46.5% for BioBERT vs. 47.2% for the CNN). These attributes are characterized by a higher contextual dependency and terminological variability, both of which affect the extraction performance. The lower recall observed for the "Class Component" is particularly relevant, as it reflects the fact that both models are not yet able to establish explicit relationships between the extracted components and their corresponding product codes or contextual roles. Indeed, at this stage, both models identify component-related terms within the text, but these are neither structured hierarchically nor linked to specific device instances. This behavior is consistent with an early PoC system and is expected to improve as the model and its underlying data structures are further refined.

When comparing the external test with the validation test, both models showed a general decrease in recall across most attributes, with the reduction being more pronounced for the CNN model, while BioBERT maintained a higher performance particularly for "Product code", "Manufacturer", and "Material". For the "Class Component", performance improved for the CNN model (from 24.9% to 50.0%) and remained stable for BioBERT (approximately 54.2%). In contrast, "Joint type" showed consistently high recall values in both datasets, with a slight increase for the CNN model (from 83.7% to 92.0%) and a marginal decrease for BioBERT (from 95.9% to 92.0%). Precision rates exhibited heterogeneous behavior across attributes, with BioBERT generally showing higher or more stable precision rates in the validation test compared to the CNN model. Overall, BioBERT demonstrated greater robustness across datasets, whereas the CNN model exhibited a higher variability in performance.

For the "Manufacturer" attribute, the main challenge was the high variability in how manufacturer names appear in the brochures. Names were often written in non-standard formats, and the AI models struggled to generalize when exposed to variants that differed from the exact forms seen during training. This issue was amplified by the limited number of examples available for each manufacturer and by the fact that manufacturer names frequently appear in titles or footers, where contextual information is minimal. For "Product code", the performance was strongly affected by the presence of tables. Most codes are embedded in tabular structures that undergo a significant loss of formatting during PDF-to-text conversion. Even when using dedicated extraction software, the table layout information is often flattened, thus depriving the model of essential structural cues. For this study, an open-source library was adopted to maintain a low-cost workflow. Naturally, a more advanced and robust extraction pipeline could have been implemented if more resources had been available. In addition, product codes vary widely in their alphanumeric composition across manufacturers, offering few semantic patterns that the model can leverage to generalize. For the "Class Component" attribute, class imbalance played a critical role. Common components such as "Stem" were well-represented in the training data and were, therefore, recognized more accurately, while rarer components were frequently misclassified due to limited exposure. For "Material", variability was driven by inconsistent nomenclature (such as Ti6Al4V, Titanium Alloy, or Ti Alloy), the extensive use of abbreviations, and the frequent placement of material descriptors inside tables or lists rather than full sentences. These factors reduced the contextual consistency and made extraction more difficult for both models.

Overall, these results highlight both the potential and the current limitations of the system. On the one hand, the dramatic reduction in processing time and the solid performance in key identification fields support the feasibility of integrating the AI model into registry workflows to accelerate device classification and dictionary updates. On the other hand, the variability observed across attributes, especially those requiring contextual

interpretation, emphasizes the need for broader datasets, improved synonym coverage, and more advanced context-aware NLP models in future iterations.

Overall, the results indicate that the AI module-PoC performs well in general information extraction and in identifying device-related data but requires further improvement to accurately handle more complex attributes that depend strongly on the context.

Strengths, Limitations, and Future Perspectives of the Study

A major strength of this study is that it has developed an AI module based entirely on open-source NLP technologies for the first time within the Italian MD registry ecosystem. This module is designed to integrate seamlessly into a national, cloud-native, microservice-based infrastructure. The adopted methodology enabled the construction of a harmonized dataset from heterogeneous sources (RIAP MD-Dictionary, NJR MD-Classification, and manufacturer technical documents), thereby replicating the complexity and variability typical of real-world data acquisition workflows.

Another strength lies in the computational efficiency of AI models (in particular, in the case of the CNN-based model), which satisfied all operational requirements of the π-RIPI platform: the use of open-source tools, the reduced computational burden, no reliance on external cloud resources, and acceptable performance levels. The modular design of the pipeline further enhances its long-term sustainability: each block (extraction, segmentation, labeling, and validation) can be upgraded independently without affecting the rest of the architecture. From an operational standpoint, the comparison between the manual and automatic extraction showed a drastic reduction in processing time, demonstrating the system's intrinsic scalability and its potential to support dynamic and timely registry updates. Moreover, the model displayed promising performance in extracting the most critical device-level identification attributes ("Manufacturer", "Joint type", and "Product code"), reinforcing the suitability of NLP for repetitive, high-volume tasks with a significant operational impact.

Despite the promising results, some limitations should be acknowledged. The dataset is still under development and does not yet include real-time data acquisition through web scraping, which limits its ability to adapt to the continuous evolution of technical product documents. Moreover, the exclusion of some class attributes limits the comprehensiveness of the classification. Additionally, the limited number of technical documents analyzed, the complexity of PDF layouts, and the variety of terminology across sources posed challenges during data extraction and labeling. The nature of the documents also introduced constraints: the technical documents, product sheets, surgical technique manuals, and regulatory documents differ markedly in structure and level of detail. Some attributes (such as materials, coatings, and class components) are often absent when the document focuses on procedural information rather than product specifications. Moreover, EMDN is almost entirely absent from commercial technical documents, which substantially limits the model's ability to learn this attribute. Ever-closer collaboration between registries, manufacturers, and experts involved in device classification activities may allow the alignment of manufacturer documentation with standardized regulatory taxonomies.

Technical limitations also emerged during the extraction phase. Multi-column layouts, merged pages, complex table structures, and text embedded within images can all hinder accurate segmentation, as they introduce nonlinear reading paths and irregular sentence boundaries. Although the integration of OCR significantly improved the system's ability to capture text contained in rasterized elements, the structural complexity of many PDFs continues to pose challenges, suggesting that further methodological refinements are required to ensure consistent and reliable parsing across a variety of document formats. Future developments may cover various topics, such as managing CAPTCHA-protected

manufacturer webpages and evaluating the impact of OCR-related distortions, document layout complexity, and PDF formatting issues on the information extraction performance. The aim is to further improve the robustness of the processing pipeline across different document sources.

One methodological point to note is that the model does not yet establish hierarchical links between class components and their corresponding product codes. At this PoC stage, the system correctly identifies the component-related terms in the text but outputs them as an unstructured list, without relational mapping. This behavior is consistent with it being an early prototype, and it is expected to improve as the module evolves and additional relational logic is introduced in future development phases.

Finally, the model must be interpreted strictly as a PoC tested in a laboratory setting. The model was trained and validated in Phase 1 on a Custom-built dataset to meet the needs required in a PoC with a limited human workforce, while, in Phase 2 and external validation, it was tested in a real-world scenario. This approach was meant to provide a more complete assessment of the implemented tool. Real-world data clearly shows more complex patterns and structures on which the model was never trained. Moreover, in Phase 2 and external validation, the classification was considered as correct only if it is also correct in the right context. Indeed, in Phase 1, the model was only tested to recall single attributes of interest, while, in Phase 2 and external validation, it was tested to recall the right attribute in the right context. Phase 2 and external validation were meant to stress the assessment of the model performance in a complex scenario, and this therefore implied decreased values in the considered metrics. However, this proved how the proposed module seems promising, but an actual tool derived from our PoC needs further focus on building the training dataset. Future steps in developing a pilot based on our proof of concept (PoC) require a greater focus on building training datasets. The full production pipeline, comprising automated web scraping, ranking, multi-layer validation workflows, and integration into the registry, is not yet operational in real-world conditions.

Future developments of the AI module-PoC will need to focus on three tightly interconnected areas: the expansion and improvement of the datasets, the refinement of the model, and the progressive implementation of the complete operational pipeline within the π-RIPI infrastructure. First, increasing both the volume and the diversity of manufacturer documents is essential to enrich the linguistic variability available during training and to populate the DBAV with a more comprehensive set of synonyms, orthographic variants, and registered trade names. This process will become even more relevant when the web-scraped content is integrated, since the textual information available on manufacturer websites typically exhibits a more fluent, narrative style compared to the highly technical structure of technical documents. A broader and more heterogeneous dataset will directly improve attribute recognition, particularly within categories that currently show lower recall due to contextual ambiguity or insufficient representation during training.

Parallel to dataset expansion and improvement, the model itself will require targeted enhancements. Refining the definitions of the various attributes and reducing the semantic overlap between classes represent foundational steps to improve classification clarity. At the same time, the incorporation of other model architectures may further strengthen the system's ability to interpret complex contexts, especially for attributes with subtle distinctions or those deeply dependent on the sentence structure. Another important development direction concerns the model's current inability to establish hierarchical associations between product codes, class components, and their corresponding attributes. Enabling such reasoning would allow the system not only to extract information, but also to reproduce the internal relational logic that characterizes the classification systems of devices in registries.

Finally, a key future milestone will be the complete deployment of the AI module-PoC in the π-RIPI production environment. This will entail carrying out the entire sequence of planned activities, from automated web scraping to ranking, multi-layer validation, administrative oversight, and the final upload of the collected information into the MD-Dictionary.

5. Conclusions

This PoC study demonstrates the feasibility of an AI-based system for the automatic extraction and classification of information on hip and knee implants using heterogeneous sources, including registry taxonomies and manufacturer documentation. Despite being developed and tested in a controlled laboratory environment, the AI module-PoC showed promising performance, particularly for core identification attributes, and achieved a dramatic reduction in processing time compared with manual extraction. These results highlight the potential of integrating NLP-driven automation within national MD registries to support timely updates, improve data completeness, and enhance post-market surveillance activities.

At the same time, the study highlights important challenges that must be addressed before real-world deployment. The terminological variability across manufacturers, incomplete documentation, the structural complexity of PDFs, and the limited availability of technical documents all affected the performance of the model for several context-dependent attributes. The model's current inability to establish hierarchical relationships among product codes, components, and their attributes also limits its applicability for full registry integration.

Nonetheless, the modular architecture, the iterative construction of the DBAV, and the planned incorporation of web scraping and more advanced NLP models provide a strong foundation for future development. With expanded and improved datasets, refined attribute definitions, improved synonym coverage, and the potential integration of other model architectures, the AI module-PoC is expected to evolve into a robust and scalable tool supporting the π-RIPI platform. Moreover, this approach may contribute to a more efficient, standardized, and data-driven ecosystem for MD surveillance at the national and international level, paving the way for future applications across additional device classes and registry domains. Ultimately, it is hoped that the results of this study might form the basis of a global search for all similar descriptive documents produced by implant manufacturers over the last two decades, and that using the same AI screening on these documents might provide many missing data points.

Supplementary Materials: The following supporting information can be downloaded at https://www.mdpi.com/article/10.3390/app162110691/s1, Table S1. Distribution of the 49 technical documents used to set up the Custom-built dataset, by manufacturer and joint; Table S2. List of all selected attributes from RIAP MD-Dictionary and NJR MD-Classification; Equations file S1. List of metrics used to assess model predictivity ability; Figure S1. Detailed description of the operational flow of the AI module-PoC once deployed; Excel S1. Selection grid for manual labeling and comparison with AI models; Excel S2. Complete confusion matrices for both validated models; Excel S3. Detailed results of the validation and external tests by analyzed technical document.

Author Contributions: Conceptualization, M.D.F., M.T., D.L.B., V.G. and F.G.; methodology: T.F., M.T., E.C. (Enrico Ciminello), E.C. (Eugenio Carrani), D.L.B., A.M., M.B. and M.Z.; software: A.M.; validation: T.F., E.C. (Enrico Ciminello), M.M., A.M. and M.T.; formal analysis, T.F., E.C. (Enrico Ciminello) and A.M.; investigation, A.M., T.F., M.B. and M.Z.; data curation, A.M.; writing—original draft preparation, T.F., E.C. (Enrico Ciminello) and M.T.; writing—review and editing, P.C., E.Y., T.W., R.A. and E.C. (Edd Caton); visualization, T.F., M.T. and E.C. (Enrico Ciminello); supervision, M.D.F.

and M.T.; project administration, P.C., M.D.F. and M.T.; funding acquisition, M.D.F., M.T. and V.G. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the European Union—NextGenerationEU, National Recovery and Resilience Plan (NRRP), Mission 4 Component 2 Investment 1.5, Project ECS 0000024 Rome Technopole—CUP I83C22001000005, Sub-project: The Artificial Intelligence Module of the π-RIPI data collection platform (RIPI-AI). The views and opinions expressed are only those of the authors and do not necessarily reflect those of the European Union or the European Commission. Neither the European Union nor the granting authority can be held responsible for them.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Classification data used in this study were provided by The National Joint Registry to the Italian National Institute of Health under license and is available by application to the NJR research@njr.org.uk. Application process and fees apply. The code developed and used in this study for data processing, model training, and analysis is proprietary and is therefore not publicly available due to commercial restrictions. Details on the software environment, model configuration, and analysis parameters are provided in Section 2.4 of the manuscript. Further requests can be directed to the corresponding author.

Acknowledgments: We acknowledge Simona Clementi, Attanasio Cornacchia, Dario Bevilacqua, Luca Pattarini, Mauro Cislaghi, and Johnny Della Giustina for their contribution to the development of the study; Francesca Capone for her role as Spoke 2 referent for the Italian National Institute of Health within the Rome Technopole project; and Alessia Biondi, Stefania Ceccarelli, and Paola Laricchiuta for their technical and scientific support to the RIPI and RIAP projects. The authors gratefully acknowledge the Master's Program in Gestione Integrata dei Dispositivi Medici-Medical Technologies Regulatory Affairs at Sapienza University of Rome for supporting the development of part of this study.

Conflicts of Interest: Alessandro Mentuccia, Marco Davide Forlani, Martina Barral, and Martina Zanforlini were employed by the BVTech company. Richard Armstrong and Edd Caton were employed by the NEC Software Solutions company. Fernanda Gellona and Valeria Glorioso were employed by the Confindustria Dispositivi Medici organization. In particular, BV Tech is one of the leading Italian players in the fields of Information and Communication Technology (ICT), cybersecurity, cloud computing, and digital transformation; NEC Software Solutions UK Ltd provides software and digital services worldwide in healthcare, housing, government and public safety and is contracted to manage and develop the core data collection, software infrastructure, and service desk for the NJR; Confindustria Dispositivi Medici is the federation within Confindustria that unites, represents, and promotes companies operating in Italy in the medical device sector. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AI	Artificial Intelligence
BioBERT	Biomedical Bidirectional Encoder Representations from Transformers for Biomedical Text Mining
CAPTCHA	Completely Automated Public Turing test to tell Computers and Humans Apart
CND	Classificazione Nazionale Dispositivi medici (medical device national classification)
CNN	Convolutional Neural Network
DBAV	DataBase of Attributes and Values
EMDN	European Medical Device Nomenclature
ETL	Extract, Transform, Load
FN	False Negative

FP	False Positive
HDR	Hospital Discharge Records
hNER	hybrid Named Entity Recognition
MD	Medical Device
NaN	Not a Number
NER	Named Entity Recognition
NJR	National Joint Registry
NLP	Natural Language Processing
OCR	Optical Character Recognition
PDF	Portable Document Format
PoC	Proof of Concept
RIAP	Registro Italiano ArtroProtesi (Italian Arthroplasty Registry)
RIPI	Registro nazionale delle protesi impiantabili (Italian Implantable Prostheses Registry)
STROBE	STrengthening the Reporting of OBservational studies in Epidemiology
TN	True Negative
TP	True Positive
UK	United Kingdom

References

1. Lübbeke, A.; Silman, A.J.; Prieto-Alhambra, D.; Adler, A.I.; Barea, C.; Carr, A.J. The role of national registries in improving patient safety for hip and knee replacements. *BMC Musculoskelet. Disord.* **2017**, *18*, 414. [CrossRef]
2. French, J.M.R.; Deere, K.; Whitehouse, M.R.; Pegg, D.J.; Ciminello, E.; Valentini, R.; Torre, M.; Mäkelä, K.; Lübbeke, A.; Bohm, E.R.; et al. The completeness of national hip and knee replacement registers. *Acta Orthop.* **2024**, *95*, 654–660. [CrossRef]
3. EUR-Lex. Access to European Union Law. Regulation (EU) 2017/745 of the European Parliament and of the Council of 5 April 2017 on Medical Devices, Amending Directive 2001/83/EC, Regulation (EC) No 178/2002 and Regulation (EC) No 1223/2009 and Repealing Council Directives 90/385/EEC and 93/42/EEC. Available online: https://eur-lex.europa.eu/eli/reg/2017/745/oj/eng (accessed on 21 July 2025).
4. Melvin, T.; Torre, M. New medical device regulations: The regulator's view. *EFORT Open Rev.* **2019**, *4*, 351–356. [CrossRef]
5. Varnum, C.; Bečić Pedersen, A.; Rolfson, O.; Rogmark, C.; Furnes, O.; Hallan, G.; Mäkelä, K.; de Steiger, R.; Porter, M.; Overgaard, S. Impact of hip arthroplasty registers on orthopaedic practice and perspectives for the future. *EFORT Open Rev.* **2019**, *4*, 368–376. [CrossRef]
6. Gerbers, J.G.; Nelissen, R.G.H.H. Arthroplasty registries: A valuable tool to orthopaedic surgeons, manufacturers, patients and regulators? *Orthop. Trauma* **2024**, *38*, 118–123. [CrossRef]
7. Graves, S.E. The value of arthroplasty registry data. *Acta Orthop.* **2010**, *81*, 8–9. [CrossRef]
8. Malchau, H.; Graves, S.E.; Porter, M.; Malchau, H.; Graves, S.E.; Porter, M.; Harris, W.H.; Troelsen, A. The next critical role of orthopedic registries. *Acta Orthop.* **2015**, *86*, 3–4. [CrossRef]
9. Lübbeke, A.; Silman, A.J.; Barea, C.; Prieto-Alhambra, D.; Carr, A.J. Mapping existing hip and knee replacement registries in Europe. *Health Policy* **2018**, *122*, 548–557. [CrossRef]
10. Ciminello, E.; Cuccu, A.; Romanini, E.; Venosa, M.; Cazzato, G.; Tucci, G.; Boniforti, F.; Carpanese, L.; Falcone, T.; Ceccarelli, S.; et al. Total hip arthroplasty in Italy: An observational, population-based study on surgical volume growth from 2001 to 2023 and forecasts until 2050 with six different statistical models. *J. Orthop. Traumatol.* **2026**, *27*, 2. [CrossRef]
11. Varnum, C.; Pedersen, A.B.; Gundtoft, P.H.; Overgaard, S. The what, when and how of orthopaedic registers: An introduction into register-based research. *EFORT Open Rev.* **2019**, *4*, 337–343. [CrossRef]
12. Toccaceli, V.; Torre, M. A far-reaching Regulation for the Italian National Registry of Implantable Prostheses: A possible model for other health registries. *Ann. Ist. Super. Sanita* **2021**, *57*, 132–137. [CrossRef]
13. Denissen, G.A.W.; van Steenbergen, L.N.; Lollinga, W.T.; Verdonschot, N.J.J.; Schreurs, B.W.; Nelissen, R.G.H.H. Generic implant classification enables comparison across implant designs: The Dutch Arthroplasty Register implant library. *EFORT Open Rev.* **2019**, *4*, 344–350. [CrossRef]
14. Endoprothesenregister Deutschland (EPRD). The German Arthroplasty Registry (EPRD) 2023. Available online: https://www.eprd.de/fileadmin/user_upload/Dateien/Publikationen/Artikel_und_Aufsaetze/EPRD_Profile_v3.0_2023-12-06_E.pdf (accessed on 21 July 2026).

15. Jansson, V.; Grimberg, A.; Melsheimer, O.; Perka, C.; Steinbrück, A. Orthopaedic registries: The German experience. *EFORT Open Rev.* **2019**, *4*, 401–408. [CrossRef]
16. Etkin, C.D.; Springer, B.D. The American Joint Replacement Registry: The first 5 years. *Arthroplast Today* **2017**, *3*, 67–69. [CrossRef]
17. Mari, V.; Pascucci, S.; Piazzolla, A.; Berjano, P.; Franzò, M.; Sampaolo, L.; Carrani, E.; Torre, M. Developing a Classification of Spinal Medical Devices: Has the Time Come? Review of the Literature and a Proposal for Spine Registries. *Bioengineering* **2025**, *12*, 853. [CrossRef]
18. Nickerson, R.C.; Varshney, U.; Muntermann, J. A method for taxonomy development and its application in information systems. *Eur. J. Inf. Syst.* **2013**, *22*, 336–359. [CrossRef]
19. Sasanelli, F.; Le, K.D.R.; Tay, S.B.P.; Tran, P.; Verjans, J.W. Applications of Natural Language Processing Tools in Orthopaedic Surgery: A Scoping Review. *Appl. Sci.* **2023**, *13*, 11586. [CrossRef]
20. Liu, P.; Zhang, J.; Liu, S.; Huo, T.; He, J.; Xue, M.; Fang, Y.; Wang, H.; Xie, Y.; Xie, M.; et al. Application of artificial intelligence technology in the field of orthopedics: A narrative review. *Artif. Intell. Rev.* **2024**, *57*, 13. [CrossRef]
21. Aydın, M.; Orhan, F. Evaluating the Impact of AI in Orthopedics: A Quantitative Analysis of Advancements and Challenges. *Bratisl. Med. J.* **2025**, *126*, 1858–1875. [CrossRef]
22. Akbarian, E.; Mohammadi, M.; Tiala, E.; Ljungberg, O.; Sharif Razavian, A.; Magnéli, M.; Gordon, M. Development and validation of an artificial intelligence model for the classification of hip fractures using the AO-OTA framework. *Acta Orthop.* **2024**, *95*, 340–347. [CrossRef]
23. Farhadi, F.; Barnes, M.R.; Sugito, H.R.; Sin, J.M.; Henderson, E.R.; Levy, J.J. Applications of artificial intelligence in orthopaedic surgery. *Front. Med. Technol.* **2022**, *4*, 995526. [CrossRef]
24. Geda, M.W.; Tang, Y.M.; Lee, C.K.M. Applications of artificial intelligence in orthopaedic surgery: A systematic review and meta-analysis. *Eng. Appl. Artif. Intell.* **2024**, *133*, 108326. [CrossRef]
25. Han, F.; Huang, X.; Wang, X.; Chen, Y.F.; Lu, C.; Li, S.; Lu, L.; Zhang, D.W. Artificial Intelligence in Orthopedic Surgery: Current Applications, Challenges, and Future Directions. *MedComm* **2025**, *6*, e70260. [CrossRef]
26. Ren, Y.; Caiani, E.G. Leveraging natural language processing to aggregate field safety notices of medical devices across the EU. *npj Digit. Med.* **2024**, *7*, 352. [CrossRef]
27. Confindustria Dispositivi Medici. Available online: https://www.confindustriadm.it/ (accessed on 21 July 2026).
28. BVTech. Available online: https://www.bvtech.com/en/ (accessed on 21 July 2026).
29. Italian Implantable Prostheses Registry. Available online: https://ripi.iss.it/ripi/en/p-ripi/ (accessed on 21 July 2026).
30. Bacocco, D.; Carrani, E.; Ciciani, B.; Di Sanzo, P.; Leotta, F.; Torre, M. Design and implementation of the new Italian healthcare digital interoperable registry for implantable medical devices. *Softw. Pract. Exp.* **2022**, *52*, 2368–2392. [CrossRef]
31. von Elm, E.; Altman, D.G.; Egger, M.; Pocock, S.J.; Gøtzsche, P.C.; Vandenbroucke, J.P. STROBE Initiative. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: Guidelines for reporting observational studies. *J. Clin. Epidemiol.* **2008**, *61*, 344–349. [CrossRef]
32. Ramachandran, R.; Arutchelvan, K. Named entity recognition on bio-medical literature documents using hybrid based approach. *J. Ambient Intell. Hum. Comput* **2021**, *11*, 1–10. [CrossRef]
33. SpaCy. Available online: https://spacy.io/models/en/#en_core_web_sm (accessed on 21 July 2026).
34. BioBERT. Available online: https://huggingface.co/dmis-lab/biobert-base-cased-v1.1 (accessed on 21 July 2026).
35. Github. Available online: https://github.com/dmis-lab/biobert (accessed on 21 July 2026).
36. Pruneski, J.A.; Pareek, A.; Nwachukwu, B.U.; Martin, R.K.; Kelly, B.T.; Karlsson, J.; Pearle, A.D.; Kiapour, A.M.; Williams, R.J., 3rd. Natural language processing: Using artificial intelligence to understand human language in orthopedics. *Knee Surg. Sports Traumatol. Arthrosc.* **2023**, *31*, 1203–1211. [CrossRef]
37. Karnuta, J.M.; Haeberle, H.S.; Luu, B.C.; Roth, A.L.; Molloy, R.M.; Nystrom, L.M.; Piuzzi, N.S.; Jonathan, L.; Schaffer, J.L.; Chen, A.F.; et al. Artificial Intelligence to Identify Arthroplasty Implants From Radiographs of the Hip. *J. Arthroplast.* **2021**, *36*, S290–S294. [CrossRef]
38. Ansoborlo, M.; Cardoso, J.; Herbert, J.; Salpetrier, C.; Bouzillé, G.; Cuggia, M.; Rosset, P.; Le Nail, L.R.; Grammatico-Guillon, L. Performance of a NLP Tool for Text Classification from Orthopaedic Operative Reports, Using Data from the Large Network of Clinical Data Warehouses of the West of France: The HACROHUGORTHO Project. *Stud. Health Technol. Inf.* **2024**, *316*, 1979–1983. [CrossRef]
39. Ansoborlo, M.; Salpétrier, C.; Nail, L.L.; Herbet, J.; Cuggia, M.; Rosset, P.; Grammatico-Guillon, L. Feasibility of automated surveillance of implantable devices in orthopaedics via clinical data warehouse: The Studio study. *BMC Med. Inf. Decis. Mak.* **2024**, *24*, 324. [CrossRef]
40. European Commission—Public Health. European Medical Devices Nomenclature (EMDN). Available online: https://health.ec.europa.eu/medical-devices-topics-interest/european-medical-devices-nomenclature-emdn_en (accessed on 21 July 2026).

41. Jolly, A.; Pandey, V.; Singh, I.; Sharma, N. Exploring Biomedical Named Entity Recognition via SciSpaCy and BioBERT Models. *Open BioMed Eng. J.* **2024**, *18*, e18741207289680. [CrossRef]
42. Zhang, Y.; Qi, P.; Manning, C.D.; Langlotz, C.P. Biomedical and clinical English model packages for the Stanza Python NLP library. *J. Am. Med. Inform. Assoc.* **2021**, *28*, 1892–1899. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.