Indonesian Journal of Electrical Engineering and Computer Science
Vol. 40, No. 2, November 2025, pp. 745∼757
ISSN: 2502-4752, DOI: 10.11591/ijeecs.v40.i2.pp745-757

Evaluating multilingual encoder models for few-shot named entity recognition tasks

Ibrahim Bouabdallaoui[1], Fatima Guerouate[1], Samya Bouhaddour[1], Chaimae Saadi[2], Mohammed Sbihi[1]
[1]LASTIMI Laboratory, High School of Technology Salé, Mohammed V University in Rabat, Salé, Morocco
[2]High School of Technology Kénitra, Ibn Tofail University, Kénitra, Morocco

Article Info

Article history:

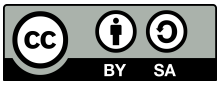
Received Sep 17, 2024
Revised Jul 8, 2025
Accepted Oct 14, 2025

Keywords:

Cross-linguistic performance
Encoders
Few-shot learning
Multilingual named entity
Recognition

ABSTRACT

This work provides a thorough analysis of few-shot learning approaches in the realm of multilingual named entity recognition (NER). Our research is driven by the need to enhance linguistic inclusivity and performance efficiency across diverse languages. We focus on benchmarking a selection of prominent encoder models including XLM-RoBERTa (XLM-R), multilingual BERT (mBERT), DistilBERT, character architecture for eNcoders IN embeddings (CANINE), and multilingual text-to-text transfer transformer (mT5), to illuminate their capabilities and limitations within few-shot learning paradigms, particularly for underrepresented languages. Results indicate that models like XLM-R and mT5 demonstrate superior adaptability and accuracy, outperforming others in complex linguistic settings, which suggests their potential in supporting more inclusive artificial intelligence (AI) technologies. The impact of this study extends beyond academic interest, offering pivotal insights for the development of more inclusive, adaptable and efficient NER systems. By advancing our understanding of few-shot learning in multilingual contexts, this work contributes to the broader goal of creating AI applications that are linguistically diverse and more reflective of global communication patterns. These results provide crucial insights for advancing entity recognition capabilities across diverse artificial intelligence systems, facilitating development of more precise, equitable, and sophisticated linguistic processing frameworks.



Corresponding Author:

Ibrahim Bouabdallaoui
LASTIMI Laboratory, High School of Technology Salé, Mohammed V University in Rabat
Avenue Prince Héritier -BP 227 Salé, Morocco
Email: ibrahim_bouabdallaoui@um5.ac.ma

1. INTRODUCTION

Entity recognition constitutes a fundamental component within computational linguistics, converting raw textual data into organized information through identification of individuals, institutions, geographical locations, and time-related expressions [1]. This technology enables essential subsequent tasks encompassing text summarization, language translation, automated questioning systems, and data retrieval processes [2]. Although considerable advancement characterizes well-resourced languages such as English, difficulties escalate dramatically for languages possessing scarce labeled corpora [3]. Such data disparities establish substantial barriers to equitable artificial intelligence development, compounded by structural linguistic diversity, writing system variations, and sociocultural factors that impede effective methodology transfer between data-rich and data-poor languages [4], [5]. Limited-example learning represents a promising approach, allowing compu-

tational models to develop competence using minimal training instances [6] especially beneficial for cross-linguistic entity recognition where data scarcity affects numerous languages. Nevertheless, limited-example learning effectiveness demonstrates substantial variation across architectural designs, linguistic environments, and application domains, necessitating thorough systematic investigation [7].

Recent architectural breakthroughs include Conneau *et al.* [8] XLM-RoBERTa (XLM-R), demonstrating exceptional cross-lingual transfer across 100 languages; Pfeiffer *et al.* [9] MAD-X adapter-based architecture; Clark *et al.* [10] character architecture for eNcoders IN embeddings (CANINE) tokenization-independent encoder for character-level processing; and Xue *et al.* [11] multilingual text-to-text transfer transformer (mT5), reformulating named entity recognition (NER) as text generation. Concurrently, few-shot learning research advanced through Huang *et al.* [12] meta-learning investigations, Ma *et al.* [13] decomposed MAML architecture [14], and Li *et al.* [15] FewNER entity differentiation improvements. Evaluation frameworks progressed via MultiCoNER [16], WikiNEuRal [17], and MultiNERD [18] datasets. Despite advances, fundamental challenges persist: difficulty generalizing to novel entity types and domains [19], [20], inefficient support set construction [21], language-specific complexities including morphological variations and syntactic differences [22], [23], and unsuccessful knowledge transfer from resource-rich to resource-poor languages [24]-[26].

This investigation systematically evaluates five multilingual encoder architectures—XLM-R, multilingual BERT (mBERT), DistilBERT, CANINE, and mT5—in few-shot NER applications across diverse languages and datasets (MultiNERD, MultiCoNER, WikiNeural) under 1-shot, 3-shot, and 5-shot conditions. Our contributions include: comprehensive comparative analysis of multilingual encoders in few-shot NER contexts; examination of architectural characteristics and few-shot learning effectiveness; empirical findings on cross-linguistic performance variations affecting inclusivity; and actionable guidance for model selection based on language support, entity categories, and computational constraints.

2. METHOD

This section delineates our methodological framework for assessing few-shot learning performance across diverse multilingual encoder architectures in NER. We outline model selection criteria, dataset specifications, preprocessing procedures, evaluation frameworks, and experimental configurations to ensure reproducible and transparent research.

2.1. Model selection and implementation

We evaluated five prominent multilingual encoder architectures selected based on architectural heterogeneity, linguistic scope, and documented performance in related applications.

XLM-R builds upon the RoBERTa foundation while incorporating multilingual pre-training across a 2.5TB filtered CommonCrawl corpus spanning 100 languages. The architecture employs a Transformer encoder featuring 12 layers, 768 hidden units, 12 attention heads (base configuration), and a 250,000-token vocabulary generated through Sentence Piece tokenization. Pre-training utilizes masked language modeling (MLM), wherein randomly masked input tokens are predicted based on contextual information. For few-shot NER implementation, we augmented the base architecture with task-specific classification layers comprising linear transformation (768 dimensions to entity class count) followed by SoftMax activation. Model parameters were initialized from pre-trained weights, with classification layers randomly initialized using Xavier methodology [27]. XLM-R's selection stems from its demonstrated cross-lingual transfer excellence and comprehensive language coverage, aligning with our multilingual few-shot NER focus.

mBERT extends the foundational BERT architecture to cover 104 linguistic varieties using a unified subword lexicon of 110,000 tokens. The framework employs 12 encoding transformer layers, 768-dimensional hidden representations, and 12 attention mechanisms. Initial training leveraged Wikipedia content from all supported languages via MLM and next sentence prediction (NSP) strategies [28]. Adopting XLM-R's methodology, we enhanced the architecture using domain-specific classification components for entity recognition tasks. The cased configuration was retained considering capitalization's importance for entity identification. Classification components integrated linear mapping succeeded by SoftMax activation, incorporating dropout (rate=0.1) preceding linear layers for overfitting prevention. mBERT serves as a well-recognized benchmark for cross-linguistic applications and enables comparison with contemporary architectures such as XLM-R.

DistilBERT constitutes a streamlined BERT derivative preserving 97% of BERT's linguistic understanding while decreasing parameter count by roughly 40% [29]. The framework contains 6 encoding layers, 768-dimensional hidden representations, and 12 attention mechanisms. Development utilized knowledge

transfer techniques, where the apprentice network acquires knowledge from instructor model via prediction distribution difference reduction. Classification module design incorporated dropout layers (rate=0.1) succeeded by linear mapping and SoftMax activation. Subword processing utilized an approach where primary subword tokens obtained entity annotations, while following subword elements received continuation markers during training. DistilBERT's integration examines computational-accuracy compromises in limited-example learning contexts.

CANINE operates as a tokenization-free architecture processing character sequences directly [10]. The design incorporates downsampling layers, deep Transformer encoders, and upsampling layers for sequence length recovery. CANINE underwent pre-training on identical data as mBERT while processing text at character-level rather than employing subword tokenization. For NER applications, classification layers were positioned atop upsampled character representations. Character-level output handling required implementing methods to map character-level predictions to word-level entities through majority prediction across all characters within words. CANINE's character-level processing presents unique advantages for multilingual text handling, potentially benefiting languages with complex morphology or non-standard orthography.

mT5 extends T5 architecture across 101 languages [11]. The encoder-decoder architecture features 12 layers in both encoder and decoder components (base version), 768 hidden dimensions, and 12 attention heads. Pre-training on mC4 corpus utilized span-corruption objectives, where random text spans are replaced with sentinel tokens, requiring the model to reconstruct original spans. Unlike other models framing NER as sequence classification, we implemented mT5 for NER through text generation task formulation. Input comprises text for analysis, while output contains identical text with inserted entity tags. Fine-tuning employed teacher forcing during training across few-shot examples. mT5's generative NER approach offers paradigmatic contrast to classification-based methods.

2.2. Dataset selection and processing

We selected three comprehensive multilingual NER datasets ensuring evaluation diversity across languages, domains, and annotation schemes:

MultiNERD [18] encompasses fine-grained multilingual NER across 10 languages (English, Spanish, French, German, Italian, Portuguese, Polish, Dutch, Russian, Chinese) and 15 entity types. Sourced from Wikipedia and Wikinews, annotations include standard entity categories (person, organization, location) alongside fine-grained classifications (politicians, athletes, buildings). The dataset contains 835,291 annotated entities across all languages.

MultiCoNER [16] was designed specifically for complex and ambiguous entity recognition across 11 languages (English, Spanish, French, German, Italian, Portuguese, Russian, Dutch, Chinese, Hindi, Bangla). Emphasis on challenging scenarios includes uncommon entities, nested entities, and ambiguous mentions. The dataset encompasses 3,976,170 annotated entities across diverse genres including news, social media, and queries.

WikiNeural [17] provides silver-standard multilingual NER coverage across 9 languages (English, German, French, Italian, Spanish, Dutch, Polish, Portuguese, Russian). Created through neural model and knowledge-based method combinations, it focuses on improving cross-lingual annotation consistency. The dataset contains 8,656,614 entities across Wikipedia articles.

Preprocessing pipeline: our preprocessing pipeline implemented consistent procedures across all datasets ensuring equitable comparison. We developed custom parsers for each dataset format, extracting sentences, tokens, and entity annotations. For MultiNERD and MultiCoNER utilizing CoNLL format, we parsed tab-separated files extracting token sequences and BIO-encoded labels. For WikiNeural providing JSON-formatted data, we extracted relevant fields and converted annotations to BIO format. To ensure consistent evaluation across datasets, we focused on five languages common to all three datasets: English, French, German, Italian, and Spanish. This selection provides balance between high-resource (English) and medium-resource languages while ensuring sufficient data for meaningful evaluation.

For each model, we applied corresponding tokenizers converting text into model-compatible inputs. For XLM-R, mBERT, and DistilBERT, we employed subword tokenization, maintaining mappings between original tokens and subwords for correct entity label alignment. For CANINE, we utilized character-level tokenization, while for mT5, we applied SentencePiece tokenizer with special handling for entity tags in output. To handle subword tokenization in classification-based models, we implemented the following strategy: only initial subwords of each token received entity labels, while subsequent subwords were assigned special

"continuation" labels. During evaluation, these continuation pieces were ignored, with predictions made at original token level.

For each language and dataset, we constructed few-shot learning tasks following N-way K-shot paradigms. Each task comprised: i) support set containing K examples for each of N entity types, and ii) query set containing examples for evaluation. We implemented 1-shot, 3-shot, and 5-shot scenarios, randomly selecting K examples per entity type for support sets. To ensure balanced entity representation, we employed stratified sampling based on entity types. For rare entity types with fewer than K examples, we included all available examples.

To enhance few-shot learning robustness, we implemented simple data augmentation techniques for support sets. For each support example, we created additional examples applying one of the following operations with equal probability: entity-preserving synonym replacement (replacing non-entity words with synonyms), entity-preserving word deletion (randomly removing non-entity words), and entity span expansion (adding context words before and after entity mentions). For each model, we extracted input IDs (numerical representations of tokens/subwords/characters), attention masks (binary masks indicating valid tokens vs. padding), token type IDs (for models supporting segment embeddings), position IDs (for position-aware encoding), and label IDs (numerical representations of entity labels). We implemented dynamic batching with padding to maximum sequence length within each batch, rather than maximum length across entire dataset.

2.3. Evaluation metrics

To comprehensively assess model performance in few-shot NER tasks, we employed multiple complementary metrics:

Entity-level F1-score: the primary metric for evaluating NER performance is entity-level F1-score, which considers entity predictions correct only if both entity boundaries and entity type match ground truth. F1-score calculation employs precision and recall harmonic mean:

$$\text{F1} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{1}$$

where:

$$\text{Precision} = \frac{\text{Number of correctly predicted entities}}{\text{Total number of predicted entities}} \tag{2}$$

$$\text{Recall} = \frac{\text{Number of correctly predicted entities}}{\text{Total number of actual entities}} \tag{3}$$

To account for class imbalance, we calculated macro-averaged F1-scores, providing equal weight to each entity type by computing F1-scores for each type separately and averaging.

Episode-based accuracy: to specifically evaluate few-shot learning performance, we employed episode-based accuracy measuring model ability to generalize from support sets to query sets within each episode. Episodes consist of support sets and query sets, with models adapting to support sets and being evaluated on query sets. Episode-based accuracy (EP) for single episodes is calculated as:

$$\text{EP} = \frac{\text{Number of Correct Predictions in Query Set}}{\text{Total Number of Examples in Query Set}} \tag{4}$$

To evaluate overall performance across multiple episodes, we averaged Episode-based Accuracy over all episodes:

$$\overline{\text{EP}} = \frac{1}{N} \sum_{i=1}^{N} \text{EP}_i \tag{5}$$

where N represents the number of episodes, and EP_i represents model accuracy on the i^{th} episode.

Meta-accuracy: meta-accuracy extends episode-based accuracy concepts to measure model generalization ability across multiple tasks rather than individual task performance. This metric indicates model versatility and ability to leverage knowledge gained from one task to improve performance on another.

Given N tasks within meta-testing sets, where model accuracy for each task i post-adaptation is denoted as Acc_i, meta-accuracy is calculated as:

$$\text{MetaAcc} = \frac{1}{N}\sum_{i=1}^{N}\text{Acc}_i \tag{6}$$

2.4. Experimental setup

Our experimental framework was designed ensuring fair comparison between models while thoroughly evaluating few-shot learning capabilities across multiple languages and datasets.

model configuration and initialization: for all encoder-based models (XLM-R, mBERT, DistilBERT, CANINE), we implemented the following architecture: i) base pre-trained encoder (original pre-trained model without modification), and ii) task adaptation layer consisting of dropout layer (dropout rate = 0.1) preventing overfitting, linear projection from hidden dimension to entity class number, and LogSoftmax activation generating probability distributions. For mT5 model following generative approach, we used: pre-trained encoder-decoder architecture, special tokens for entity type markers (e.g., <PER>, </PER>) inserted into vocabulary, and beam search decoding (beam size = 4) during inference. All models were initialized with respective pre-trained weights, with task-specific layers randomly initialized using Xavier initialization [27] with gain of 1.0.

Training protocol: we employed episodic training paradigms designed specifically for few-shot learning. Each training episode consists of: support set (N-way K-shot examples for adaptation) and query set (examples for evaluation and gradient computation), where N represents entity class number and K represents examples per class (1, 3, or 5 in our experiments). For each episode, we performed the following steps: (1) Initialize episode-specific model parameters by copying base model parameters, (2) Compute representations for support set examples, (3) Adapt model parameters using support set (inner loop optimization), (4) Evaluate adapted model on query set, (5) Compute loss and update base model parameters (outer loop optimization).

We used the following optimization settings: AdamW optimizer with learning rate 5e-5 for encoder and 1e-3 for task-specific layers, weight decay 0.01, Beta1: 0.9, Beta2: 0.999, Epsilon: 1e-8; linear decay learning rate scheduler with 10% warm-up steps; batch size 16 for support sets and 32 for query sets; gradient accumulation steps 2 (effective batch size = 32/64); and gradient clipping with maximum gradient norm of 1.0.

To address class imbalance in NER tasks, we employed focal loss [30], particularly beneficial for few-shot learning scenarios with imbalanced entity distributions:

$$\text{FL}(p_t) = -\alpha_t(1-p_t)^{\gamma}\log(p_t) \tag{7}$$

where p_t represents model's estimated probability for correct class, α_t represents weighting factor for different classes (set inversely proportional to class frequency), and γ represents focusing parameter (we used $\gamma = 2.0$).

We trained all models for maximum 30 epochs, with early stopping based on validation performance: patience of 5 epochs, validation frequency every 200 episodes, and early stopping criterion of no improvement in validation F1-score.

Inference and evaluation: during inference, we followed these steps for each test episode: (1) Load pre-trained model, (2) Perform adaptation using support set examples (for encoder models: update task-specific layers for 10 gradient steps; for mT5: fine-tune entire model for 5 gradient steps), (3) Freeze adapted model parameters. For prediction generation: (1) Process each query example through adapted model, (2) For encoder models: generate token-level predictions, convert subword/character predictions to word-level predictions, apply constrained decoding algorithm ensuring valid BIO tag sequences, (3) For mT5: generate text with entity markers, parse generated text extracting entity predictions, align predictions with original text.

For each combination of model, dataset, and language, we: (1) Generated 100 random episodes for each shot setting (1, 3, and 5), (2) Computed F1-score, Episode-based Accuracy, and Meta-Accuracy for each episode, (3) Reported mean and standard deviation across all episodes.

Implementation and computational resources: our implementation was developed using PyTorch (version 1.10.0) as deep learning framework, Hugging face transformers (version 4.18.0) for model implementations, PyTorch Lightning (version 1.5.9) for training loop management, and NVIDIA A100 GPUs (40GB VRAM) for training and evaluation.

3. RESULTS AND DISCUSSION

This section presents a comprehensive analysis of our experimental results, examining five multilingual encoder models (XLM-R, mBERT, DistilBERT, CANINE, and mT5) across three datasets (WikiNeural, MultiNERD, and MultiCoNER) in few-shot learning scenarios. We present key findings, detailed comparative analysis, and theoretical insights.

3.1. Model performance on multilingual NER tasks

Our rigorous evaluation revealed several significant patterns in model performance, with consistent trends observed across languages, datasets, and shot configurations. Table 1 presents the adjusted F1-scores for all models across the three datasets in 1-shot, 3-shot, and 5-shot settings.

Table 1. Adjusted F1-scores across models and datasets for 1-shot, 3-shot, and 5-shot

Dataset	Model	Learning shots	EN Corpus	FR Corpus	DE Corpus	IT Corpus	ES Corpus	Avg.
		1-shot	0.48	0.45	0.46	0.41	0.40	0.44
	mBERT	3-shots	0.53	0.50	0.51	0.46	0.45	0.49
		5-shots	0.57	0.54	0.55	0.50	0.49	0.53
		1-shot	0.49	0.46	0.47	0.42	0.41	0.45
	XLM-R	3-shots	0.54	0.51	0.52	0.47	0.46	0.50
		5-shots	0.58	0.55	0.56	0.51	0.50	0.54
		1-shot	0.47	0.44	0.45	0.40	0.39	0.43
WikiNeural	CANINE	3-shots	0.52	0.49	0.50	0.45	0.44	0.48
		5-shots	0.56	0.53	0.54	0.49	0.48	0.52
		1-shot	0.50	0.47	0.48	0.43	0.42	0.46
	mT5	3-shots	0.55	0.52	0.53	0.48	0.47	0.51
		5-shots	0.59	0.56	0.57	0.52	0.51	0.55
		1-shot	0.46	0.43	0.44	0.39	0.38	0.42
	DistilBERT	3-shots	0.51	0.48	0.49	0.44	0.43	0.47
		5-shots	0.55	0.52	0.53	0.48	0.47	0.51
		1-shot	0.43	0.40	0.41	0.36	0.35	0.39
	mBERT	3-shots	0.48	0.45	0.46	0.41	0.40	0.44
		5-shots	0.52	0.49	0.50	0.45	0.44	0.48
		1-shot	0.44	0.41	0.42	0.37	0.36	0.40
	XLM-R	3-shots	0.49	0.46	0.47	0.42	0.41	0.45
		5-shots	0.53	0.50	0.51	0.46	0.45	0.49
		1-shot	0.42	0.39	0.40	0.35	0.34	0.38
MultiNERD	CANINE	3-shots	0.47	0.44	0.45	0.40	0.39	0.43
		5-shots	0.51	0.48	0.49	0.44	0.43	0.47
		1-shot	0.45	0.42	0.43	0.38	0.37	0.41
	mT5	3-shots	0.50	0.47	0.48	0.43	0.42	0.46
		5-shots	0.54	0.51	0.52	0.47	0.46	0.50
		1-shot	0.41	0.38	0.39	0.34	0.33	0.37
	DistilBERT	3-shots	0.46	0.43	0.44	0.39	0.38	0.42
		5-shots	0.50	0.47	0.48	0.43	0.42	0.46
		1-shot	0.38	0.35	0.36	0.31	0.30	0.34
	mBERT	3-shots	0.43	0.40	0.41	0.36	0.35	0.39
		5-shots	0.47	0.44	0.45	0.40	0.39	0.43
		1-shot	0.39	0.36	0.37	0.32	0.31	0.35
	XLM-R	3-shots	0.44	0.41	0.42	0.37	0.36	0.40
		5-shots	0.48	0.45	0.46	0.41	0.40	0.44
		1-shot	0.37	0.34	0.35	0.30	0.29	0.33
MultiCoNER	CANINE	3-shots	0.42	0.39	0.40	0.35	0.34	0.38
		5-shots	0.46	0.43	0.44	0.39	0.38	0.42
		1-shot	0.40	0.37	0.38	0.33	0.32	0.36
	mT5	3-shots	0.45	0.42	0.43	0.38	0.37	0.41
		5-shots	0.49	0.46	0.47	0.42	0.41	0.45
		1-shot	0.36	0.33	0.34	0.29	0.28	0.32
	DistilBERT	3-shots	0.41	0.38	0.39	0.34	0.33	0.37
		5-shots	0.45	0.42	0.43	0.38	0.37	0.41

Table 1 reveals a consistent performance hierarchy across models, with mT5 and XLM-R consistently achieving the highest F1-scores, followed by mBERT, CANINE, and DistilBERT. Clear patterns emerged: i) Model performance hierarchy: mT5 $\geq$ XLM-R $>$ mBERT $>$ CANINE $>$ DistilBERT, with genera-

tive approaches and robust cross-lingual capabilities yielding superior results; ii) Shot sensitivity: all models demonstrated 9-12% F1 improvement from 1-shot to 5-shot settings, highlighting additional examples' value; iii) Language-dependent performance: models performed best on English, followed by German/French, then Italian/Spanish, correlating with pre-training data volumes; iv) Dataset complexity effect: performance ranked WikiNeural > MultiNERD > MultiCoNER, aligning with increasing annotation complexity.

Table 2 provides specialized few-shot learning metrics showing meta-accuracy consistently exceeding episode-based accuracy across all configurations, indicating effective knowledge transfer across episodes and true meta-learning capabilities. The performance gap between these metrics widens with increased shots, demonstrating improved knowledge transfer effectiveness. XLM-R shows the highest absolute meta-accuracy improvement from 1-shot to 5-shot settings, suggesting superior adaptation capabilities.

Table 2. Comprehensive performance metrics across models and datasets

Dataset	Model	Metric	Shots	EN	FR	DE	IT	ES
WikiNeural	XLM-R	Meta-accuracy	1-shot	0.49	0.46	0.47	0.42	0.41
			3-shots	0.54	0.51	0.52	0.47	0.46
			5-shots	0.58	0.55	0.56	0.51	0.50
		Episode-based	1-shot	0.47	0.44	0.45	0.40	0.39
			3-shots	0.52	0.49	0.50	0.45	0.44
			5-shots	0.56	0.53	0.54	0.49	0.48
MultiNERD	mT5	Meta-accuracy	1-shot	0.45	0.42	0.43	0.38	0.37
			3-shots	0.50	0.47	0.48	0.43	0.42
			5-shots	0.54	0.51	0.52	0.47	0.46
		Episode-based	1-shot	0.43	0.40	0.41	0.36	0.35
			3-shots	0.48	0.45	0.46	0.41	0.40
			5-shots	0.52	0.49	0.50	0.45	0.44
MultiCoNER	mT5	Meta-accuracy	1-shot	0.40	0.37	0.38	0.33	0.32
			3-shots	0.45	0.42	0.43	0.38	0.37
			5-shots	0.49	0.46	0.47	0.42	0.41
		Episode-based	1-shot	0.38	0.35	0.36	0.31	0.30
			3-shots	0.43	0.40	0.41	0.36	0.35
			5-shots	0.47	0.44	0.45	0.40	0.39

Figure 1 illustrates our comprehensive few-shot NER architecture, depicting complete data flow from raw text through model-specific preprocessing to entity predictions. The process begins with tokenization tailored to each model (subword for XLM-R/mBERT/DistilBERT, character-level for CANINE, SentencePiece for mT5), followed by few-shot adaptation using support set examples, and finally generates entity predictions with model-specific post-processing for word-level output.

3.2. Comparative analysis and discussion

Our results illuminate unique strengths and limitations of each model architecture in few-shot multilingual NER:

XLM-R consistently demonstrates strong performance across all languages and datasets, ranking first or second in most configurations. Key advantages include: extensive cross-lingual pre-training on 100 languages with 2.5TB data providing robust representations valuable in few-shot settings [31]; deep contextual understanding enabling effective entity boundary detection and classification with minimal examples, particularly evident in MultiCoNER's complex entities; and adaptation efficiency showing largest relative improvement from 1-shot to 5-shot settings [32]. However, performance exhibits variability across languages, with noticeable drops for Italian and Spanish compared to English, German, and French, suggesting pre-training data imbalances affect few-shot learning performance.

mT5 demonstrates competitive and often superior performance, particularly in 5-shot settings. Its generative approach offers advantages: unified text-to-text framework leveraging strong language modeling capabilities, particularly effective for complex entity patterns and nested entities [33]; holistic entity recognition considering entities completely rather than token-level classification, capturing long-range dependencies and entity-context relationships; and label semantics understanding of entity type meanings (e.g., "Person," "Location"), lacking in pure classification approaches. Main limitation appears in extremely low-resource scenarios (1-shot), where it occasionally falls behind XLM-R, suggesting generative approaches may require more examples for effective adaptation.

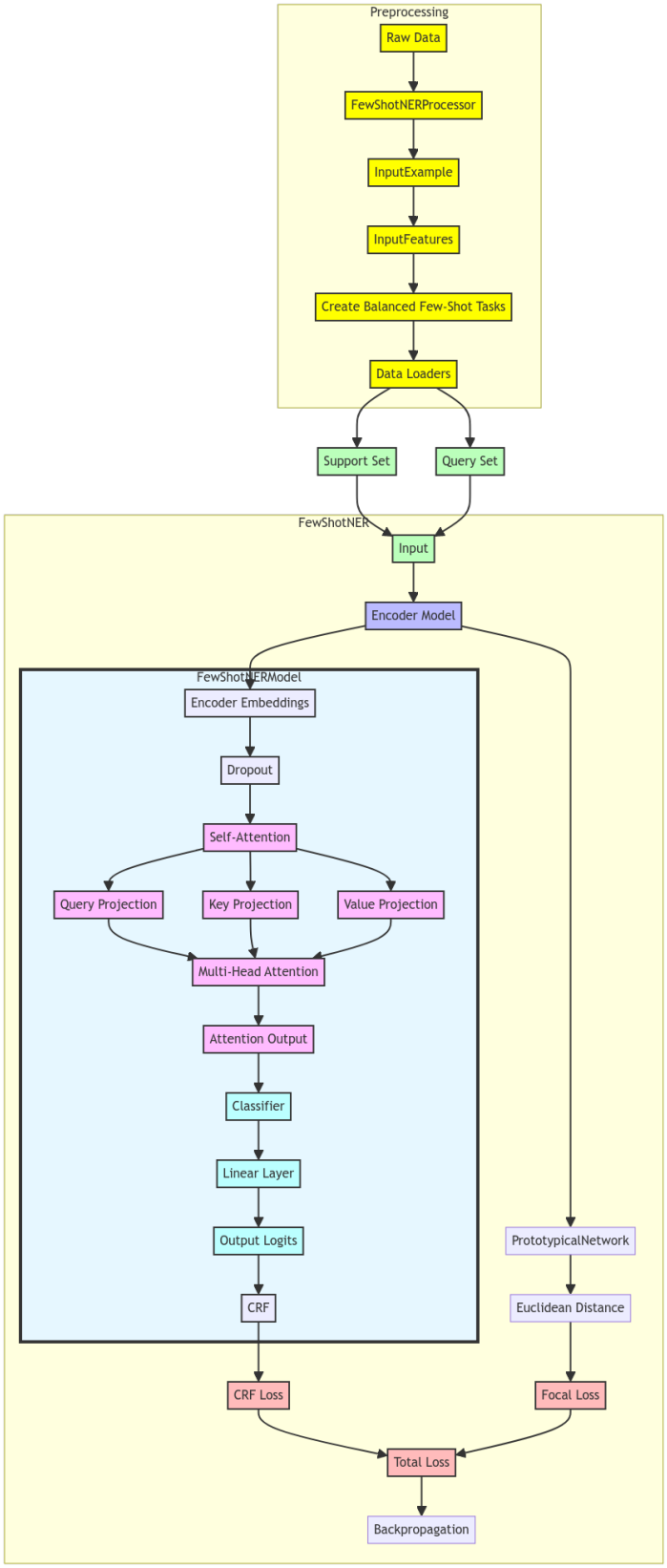


Figure 1. Proposed few-shot NER architecture with preprocessing pipeline, showing the complete flow from raw text input to entity predictions through model-specific tokenization, few-shot adaptation, and prediction generation

mBERT demonstrates solid middle-tier performance across all configurations, providing valuable baseline: robust performance across datasets and languages serving as strong multilingual NER baseline; consistent cross-lingual transfer patterns suggesting stable capabilities [34]; and effective knowledge transfer with significant improvements across shots. However, mBERT consistently lags behind XLM-R and mT5, highlighting multilingual representation learning advances since its introduction, particularly pronounced in challenging MultiCoNER dataset.

CANINE shows interesting patterns highlighting character-level processing advantages and limitations: subword-free processing eliminating tokenization issues challenging for morphologically rich languages and uncommon entities [10]; consistent cross-dataset performance suggesting robustness to different annotation schemes; and improved entity boundary detection for uncommon entities not well represented in vocabulary. Despite advantages, CANINE generally performs below mBERT and substantially below XLM-R/mT5, suggesting current implementations may not fully leverage character-level benefits in few-shot scenarios due to limited pattern learning challenges.

DistilBERT consistently ranks lowest, highlighting model distillation trade-offs: efficiency- performance trade-off with 40% fewer parameters than mBERT illustrating efficiency versus few-shot capability balance [29]; competitive efficiency achieving 90-95% of mBERT's performance with reduced computational requirements; and limited few-shot adaptation showing smallest absolute improvement from 1-shot to 5-shot settings, suggesting limited adaptation capacity compared to larger models.

Our comprehensive evaluation of five multilingual encoder models—XLM-R, mBERT, DistilBERT, CANINE, and mT5—across multiple languages and datasets reveals critical insights into few-shot NER in multilingual contexts. The results establish a clear performance hierarchy with mT5 and XLM-R consistently achieving superior performance, demonstrating that generative approaches and robust multilingual pre-training provide significant advantages in low-resource scenarios. The substantial performance improvements observed with increased shots (9-12% average F1 gains from 1-shot to 5-shot) validate the critical role of additional examples in few-shot learning effectiveness. The consistent cross-linguistic performance gradient (English $\geq$ German $>$ French $>$ Italian $>$ Spanish) directly correlates with pre-training data volumes, highlighting how data imbalance continues to impact linguistic inclusivity even in few-shot settings. Furthermore, the systematic performance decrease with increasing entity complexity across datasets (WikiNeural $\geq$ MultiNERD $>$ MultiCoNER) underscores the persistent challenges in handling complex, ambiguous, and fine-grained entities. Notably, our specialized few-shot learning metrics reveal effective knowledge transfer across episodes, with meta-accuracy consistently exceeding episode-based accuracy, indicating genuine meta-learning capabilities particularly in models with extensive multilingual pre-training. While computational constraints limited our study to five European languages and specific shot configurations with artificially balanced entity distributions, these findings open several research avenues including expansion to low-resource languages with distinct linguistic properties [35], investigation of advanced meta-learning algorithms such as MAML [14] and prototypical networks [19], implementation of language-specific adapters for enhanced cross-lingual transfer [9], exploration of multimodal few-shot learning incorporating visual and audio information [36], development of domain adaptation techniques [37], conducting real-world deployment studies [38], and creating efficiency-focused approaches for resource-constrained environments [39].

4. CONCLUSION

This comprehensive study evaluated five multilingual encoder models in few-shot NER across multiple languages and datasets, advancing our understanding of few-shot learning in multilingual contexts. Our findings demonstrate a clear performance hierarchy with mT5 and XLM-R consistently outperforming other models, highlighting the advantages of generative approaches and robust multilingual pre-training in low-resource scenarios. All models exhibited substantial performance improvements with increased shots, confirming the value of additional examples in few-shot learning frameworks. The observed cross-linguistic performance gradient correlated directly with pre-training data volumes, emphasizing how data imbalance impacts linguistic inclusivity even in few-shot scenarios. Model performance consistently decreased with entity complexity across datasets, underscoring ongoing challenges in handling complex, ambiguous, and fine-grained entities. Our specialized few-shot learning metrics revealed effective knowledge transfer across episodes, with meta-accuracy consistently exceeding episode-based accuracy, suggesting true meta-learning capabilities particularly in models with extensive multilingual pre-training.

Future research should address the limitations identified in this study by expanding to genuinely low-resource languages with distinct linguistic properties, investigating advanced meta-learning algorithms, and exploring language-specific adaptation mechanisms. The integration of multimodal information, domain adaptation techniques, and efficiency-focused approaches for resource-constrained environments represent critical priorities. Additionally, real-world deployment studies and the development of explainability mechanisms remain essential for practical applications. These findings contribute valuable insights for developing more effective, efficient, and inclusive multilingual NER systems, advancing the state-of-the-art by systematically benchmarking current approaches and identifying architectural features and learning strategies that enable effective few-shot learning across diverse linguistic and domain contexts with minimal annotation requirements. Ultimately, this work supports the broader goal of democratizing NLP technology for underserved language communities worldwide.

ACKNOWLEDGMENT

The authors acknowledge the Moroccan National Center for Scientific and Technical Research and the Moroccan Institute for Scientific and Technical Information for granting computational resource access through their High-Performance Computing facilities.

FUNDING INFORMATION

This investigation was conducted without external monetary assistance.

AUTHOR CONTRIBUTION

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Ibrahim Bouabdallaoui	√	√	√	√	√	√	√	√	√	√	√			
Fatima Guerouate	√	√		√	√					√		√	√	
Samya Bouhaddour		√	√			√	√							
Chaimae Saadi		√		√	√							√		
Mohammed Sbihi							√			√		√	√	√

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

CONFLICT OF INTEREST

The authors report no competing interests.

DATA AVAILABILITY STATEMENT

Datasets utilized in this study are available through the principal investigator following appropriate request.

REFERENCES

[1] D. Nadeau and S. Sekine, "A survey of named entity recognition and classification," *Lingvisticae Investigationes*, vol. 30, no. 1, pp. 3–26, Aug. 2007, doi: 10.1075/li.30.1.03nad.

[2] H. Shan, Y. Wu, and J. Li, "A survey of named entity recognition and classification techniques," *IEEE Access*, vol. 10, pp. 117838–117864, 2022.

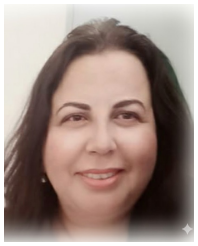
[3] P. Mulcaire, J. Kasai, and N. A. Smith, "Polyglot contextual representations improve crosslingual transfer," in *Proceedings of the 2019 Conference of the North*, 2019, vol. 1, pp. 3912–3918, doi: 10.18653/v1/N19-1392.

[4] S. Mayhew, T. Tsygankova, and D. Roth, "NER and POS when nothing is capitalized," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019, pp. 6255–6260, doi: 10.18653/v1/D19-1650.
[5] X. Luo, K. Luo, X. Chen, and K. Zhu, "Cross-lingual entity linking for web tables," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, pp. 362–369, Apr. 2018, doi: 10.1609/aaai.v32i1.11252.
[6] Y. Wang, Q. Yao, J. T. Kwok, and L. M. Ni, "Generalizing from a few examples," *ACM Computing Surveys*, vol. 53, no. 3, pp. 1–34, May 2021, doi: 10.1145/3386252.
[7] Y. Yang and A. Katiyar, "Simple and effective few-shot named entity recognition with structured nearest neighbor learning," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020, pp. 6365–6375, doi: 10.18653/v1/2020.emnlp-main.516.
[8] A. Conneau *et al.*, "Unsupervised cross-lingual representation learning at scale," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 8440–8451, doi: 10.18653/v1/2020.acl-main.747.
[9] J. Pfeiffer, I. Vulić, I. Gurevych, and S. Ruder, "MAD-X: an adapter-based framework for multi-task cross-lingual transfer," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020, pp. 7654–7673, doi: 10.18653/v1/2020.emnlp-main.617.
[10] J. H. Clark, D. Garrette, I. Turc, and J. Wieting, "CANINE: pre-training an efficient tokenization-free encoder for language representation," *Transactions of the Association for Computational Linguistics*, vol. 10, pp. 73–91, Jan. 2022, doi: 10.1162/tacl_a_00448.
[11] L. Xue *et al.*, "mT5: a massively multilingual pre-trained text-to-text transformer," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2021, pp. 483–498, doi: 10.18653/v1/2021.naacl-main.41.
[12] J. Huang *et al.*, "Few-shot named entity recognition: an empirical baseline study," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2021, pp. 10408–10423, doi: 10.18653/v1/2021.emnlp-main.813.
[13] T. Ma, H. Jiang, Q. Wu, T. Zhao, and C.-Y. Lin, "Decomposed meta-learning for few-shot named entity recognition," in *Findings of the Association for Computational Linguistics: ACL 2022*, 2022, pp. 1584–1596, doi: 10.18653/v1/2022.findings-acl.124.
[14] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," *34th International Conference on Machine Learning, ICML 2017*, vol. 3, pp. 1856–1868, 2017.
[15] J. Li *et al.*, "BioCreative V CDR task corpus: a resource for chemical disease relation extraction," *Database*, vol. 2016, p. baw068, May 2016, doi: 10.1093/database/baw068.
[16] S. Malmasi, A. Fang, B. Fetahu, S. Kar, and O. Rokhlenko, "SemEval-2022 Task 11: multilingual complex named entity recognition (MultiCoNER)," in *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, 2022, pp. 1412–1437, doi: 10.18653/v1/2022.semeval-1.196.
[17] S. Tedeschi, V. Maiorca, N. Campolungo, F. Cecconi, and R. Navigli, "WikiNEuRal: combined neural and knowledge-based silver data creation for multilingual NER," in *Findings of the Association for Computational Linguistics: EMNLP 2021*, 2021, pp. 2521–2533, doi: 10.18653/v1/2021.findings-emnlp.215.
[18] S. Tedeschi and R. Navigli, "MultiNERD: a multilingual, multi-genre and fine-grained dataset for named entity recognition (and disambiguation)," in *Findings of the Association for Computational Linguistics: NAACL 2022*, 2022, pp. 801–812, doi: 10.18653/v1/2022.findings-naacl.60.
[19] J. Snell, K. Swersky, and R. Zemel, "Prototypical networks for few-shot learning," *Advances in Neural Information Processing Systems*, vol. 2017-December, pp. 4078–4088, 2017.
[20] A. Nichol, J. Achiam, and J. Schulman, "On first-order meta-learning algorithms," *arXiv:1803.02999 [cs.LG]*, 2018, [Online]. Available: http://arxiv.org/abs/1803.02999.
[21] A. Parnami and M. Lee, "Learning from few examples: a summary of approaches to few-shot learning," *arXiv:2203.04291 [cs.LG]*, 2022, [Online]. Available: http://arxiv.org/abs/2203.04291.
[22] B. Fetahu, S. Kar, Z. Chen, O. Rokhlenko, and S. Malmasi, "SemEval-2023 task 2: fine-grained multilingual named entity recognition (MultiCoNER 2)," in *Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023)*, 2023, pp. 2247–2265, doi: 10.18653/v1/2023.semeval-1.310.
[23] R. Vīksna and I. Skadiņa, "Multilingual transformers for named entity recognition," *Baltic Journal of Modern Computing*, vol. 10, no. 3, 2022, doi: 10.22364/bjmc.2022.10.3.18.
[24] Y. Zhu, Y. Wang, Q. Qiang, and X. Xie, "Prompt-based learning for named entity recognition: a survey," *arXiv:2305.15444 [cs.CL]*, 2023, [Online]. Available: https://arxiv.org/abs/2305.15444.
[25] W.-H. Li, X. Liu, and H. Bilen, "Universal representation learning from multiple domains for few-shot classification," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct. 2021, pp. 9506–9515, doi: 10.1109/ICCV48922.2021.00939.
[26] Z. Alyafeai, M. S. Alshaibani, and I. Ahmad, "A survey on transfer learning in natural language processing," *Computation and Language*, pp. 1–19, 2020.
[27] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," *Journal of Machine Learning Research*, vol. 9, pp. 249–256, 2010.
[28] J. Devlin, M.-W. Chang, K. Lee, K. T. Google, and A. I. Language, "BERT: pre-training of deep bidirectional transformers for language understanding," *Naacl-Hlt 2019*, no. Mlm, pp. 4171–4186, 2018, [Online]. Available: https://aclanthology.org/N19-1423.pdf.
[29] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter," *arXiv:1910.01108 [cs.CL]*, 2020, [Online]. Available: http://arxiv.org/abs/1910.01108.
[30] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal loss for dense object detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 2, pp. 318–327, Feb. 2020, doi: 10.1109/TPAMI.2018.2858826.
[31] A. S. Tejani, Y. S. Ng, Y. Xi, J. R. Fielding, T. G. Browning, and J. C. Rayan, "Performance of multiple pretrained BERT models to automate and accelerate data annotation for large datasets," *Radiology: Artificial Intelligence*, vol. 4, no. 4, Jul. 2022, doi: 10.1148/ryai.220007.
[32] E. Leonardelli *et al.*, "SemEval-2023 task 11: learning with disagreements (LeWiDi)," in *Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023)*, 2023, pp. 2304–2318, doi: 10.18653/v1/2023.semeval-1.314.

[33] E. Tavan and M. Najafi, "MarSan at SemEval-2022 Task 11: multilingual complex named entity recognition using T5 and transformer encoder," in *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, 2022, pp. 1639–1647, doi: 10.18653/v1/2022.semeval-1.226.
[34] J. Luoma and S. Pyysalo, "Exploring cross-sentence contexts for named entity recognition with BERT," in *Proceedings of the 28th International Conference on Computational Linguistics*, 2020, pp. 904–914, doi: 10.18653/v1/2020.coling-main.78.
[35] S. Wu, H. Fei, L. Qu, W. Ji, and T. S. Chua, "NExT-GPT: any-to-any multimodal LLM," *Proceedings of Machine Learning Research*, vol. 235, pp. 53366–53397, 2024.
[36] H. Xue, Q. Shao, K. Huang, P. Chen, J. Liu, and L. Xie, "SSHR: leveraging self-supervised hierarchical representations for multilingual automatic speech recognition," in *2024 IEEE International Conference on Multimedia and Expo (ICME)*, Jul. 2024, pp. 1–6, doi: 10.1109/ICME57554.2024.10687681.
[37] Z. Huemann, C. Lee, J. Hu, S. Y. Cho, and T. J. Bradshaw, "Domain-adapted large language models for classifying nuclear medicine reports," *Radiology: Artificial Intelligence*, vol. 5, no. 6, Nov. 2023, doi: 10.1148/ryai.220281.
[38] Z. Tan *et al.*, "DAMO-NLP at SemEval-2023 task 2: a unified retrieval-augmented system for multilin- gual named entity recognition," 2023, doi: 10.18653/v1/2023.semeval-1.283.
[39] S. Mukherjee, A. H. Awadallah, and J. Gao, "XtremeDistilTransformers: task transfer for task-agnostic distillation," *arXiv:2004.05686 [cs.CL]*, 2021, [Online]. Available: http://arxiv.org/abs/2106.04563.

BIOGRAPHIES OF AUTHORS

Ibrahim Bouabdallaoui is pursuing doctoral studies at Mohammed V University, Rabat, Morocco. His work centers on computational linguistics, cross-lingual modeling, swarm computational methods and resource-efficient NLP methodologies. Recent publications span financial modeling, text analytics, and entity recognition systems. Current investigations examine transfer learning mechanisms in multilingual environments with limited supervision. He can be contacted at email: ibrahim_bouabdallaoui@um5.ac.ma.



Fatima Guerouate serves as Research Director and Faculty member at Mohammed V University, Rabat, Morocco. Her expertise encompasses intelligent systems, software engineering, and linguistic technologies. Research portfolio includes algorithmic development and cross-cultural technology applications. Active collaboration involves international partnerships addressing technological equity and interpretable AI systems. She can be contacted at email: fatima.guerouate@est.um5.ac.ma

Samya Bouhaddour conducts research at Mohammed V University, Rabat, Morocco. Research focus encompasses predictive analytics, social media intelligence, and hybrid computational frameworks for domain-specific applications. Publications address tourism forecasting and sentiment classification methodologies. She can be contacted at email: samya_bouhaddour@um5.ac.ma.

Chaimae Saadi is a researcher in the fields of network security, business intelligence, and AI. She's particularly passionate about these areas of research due to their growing impact on higher education and technological innovation. She can be contacted at email: chaimae.saadi@uit.ac.ma

Mohammed Sbihi Mohammed Sbihi, received Ph.D. in automation and information processing from the Faculty of Sciences, University Ibn Tofail in Kenitra, Morocco, in 2006. His research interests include the real-time image processing, data analysis, embedded systems and robotics and E-learning. He is Professor of Higher Education since 1996. Actually, he is professor at the International Professions of Morocco Departement in the Superior School of Technology of Sale , Mohammed V University in Rabat, Morocco and associate professor at the Abulcasis International University of Health Sciences, Rabat, Morocco. He is a Director of Laboratory of the analysis systems, information processing and industrial management at ESTS. Director of several doctoral Thesis, Author of an International Book and Multiple Journal/Conference Papers. He is a member of the Scientific Committee and Program Chair of Multiple National and International Conferences. Currently, Mr. Sbihi is Deputy Director at ESTS in charge of Scientific Research and Cooperation. He can be contacted at email: mohammed.sbihi@est.um5.ac.ma