Prediction of case types from non-searchable pdf documents in arabic : comparison of machine learning and deep learning with image processing

Mouad El Arrasse [1], *Youness* Khourdifi [2], *Soufyane* Mounir [1] and *Alae* El Alami [3]

[1] National School of Applied Sciences of Khouribga (Laboratory of Engineering Science and Technology), Morocco.
[2] University Sultan Moulay Slimane, Polydisciplinary Faculty of Khouribga (Laboratory of Materials Science, Mathematics and Environment), Morocco.
[3] Higher School of Technology Meknès (Laboratory of Computer Engineering and Intelligent Electrical Systems), Morocco

Abstract. The study conducted focuses on predicting the different types of judicial cases presented to Moroccan administrative courts by using court decisions in the form of non-searchable PDF documents in the Arabic language. To achieve this, we utilized image processing, text cleaning techniques, and machine learning algorithms.We carried out a comparative study using both machine learning and deep learning techniques. The experiment was conducted in two phases: first on 697 court decisions, and then on 14,207 decisions from the Administrative Court of Appeal in Marrakech. Despite the challenges associated with the Arabic language, our methods were able to efficiently extract text, leading to accurate predictions. For the experiment on 697 decisions, machine learning achieved an accuracy rate of 91%, while deep learning reached 100%. For the experiment on 14,207 decisions, machine learning obtained an accuracy of 97%, and deep learning achieved 96%.As a result, this study contributes to the existing literature on the digitization and processing of unstructured documents in the Arabic language, as well as on the prediction of judicial case types through the use of machine learning and deep learning algorithms.

Keywords :

Machine learning – Deep learning – Judicial case prediction – Non-searchable PDFs – Image processing – Text extraction.

1.Introduction :

The digitization of documents is a common and essential practice in various fields of research and application, enabling the transformation of unstructured data into manipulable structured data [1]. This transformation greatly facilitates the access, search, and manipulation of the information contained within these documents. However, a major challenge arises when dealing with non-searchable PDF documents,

particularly those generated through the digitization of paper documents [2]. These documents often contain images embedding text, making their processing and analysis difficult and labor-intensive. Furthermore, the reduced quality of scanned PDF files results in additional loss of information, which further complicates their use in research and data analysis. Thus, the need to develop efficient methods for extracting, processing, and analyzing the content of these documents has become a major concern in contemporary research [3].

Additionally, the integration of Arabic-language documents adds an extra layer of complexity due to the distinctive graphical characteristics of the language, which makes character and word recognition more challenging [4]. For instance, Arabic letters are often connected to form words, and a single letter can take on different forms depending on its position within the word, creating unique challenges for Optical Character Recognition (OCR) and textual analysis. These linguistic nuances require a sophisticated methodological approach in the development of document processing tools, highlighting the need to adopt specialized techniques to ensure precise text extraction from Arabic-language documents.

In this context, it is also crucial to provide concrete examples of processed PDF documents, along with clear details about the tools and technologies used. These elements help illustrate the practical implementation of the proposed methods and validate their applicability in real-world contexts, thereby reinforcing the impact of this study.

In this study, we aim to predict the various types of judicial cases presented before the administrative courts of appeal in Morocco, using court decisions from these courts that have been digitized from Arabic-language paper documents and treated as unstructured data [5]. To achieve this, we utilized several specialized Python libraries to clean and extract text from the documents, and we also compared the effectiveness of different machine learning algorithms in our prediction process [6].

Following this introduction, our paper will be structured into five main sections. The first section, the state of the art, will review existing research on document digitization and the analysis of Arabic-language judicial documents. Next, the methodology will detail the tools and techniques used to preprocess our dataset. The results of our study, presented in the third section, will highlight the performance of the document processing techniques employed. A discussion will then analyze these results in depth, focusing on their implications and the challenges encountered. This will be followed by a section on ethical and confidentiality considerations, addressing data protection issues and the responsible use of artificial intelligence models in a judicial context. Finally, the conclusion will summarize the key findings of our research and outline future perspectives for this field of study.

2.State of the Art :

Over the past few decades, research on the digitization and processing of non-searchable documents has seen remarkable progress, becoming a crucial field in information engineering and artificial intelligence. From the early attempts at Optical Character Recognition (OCR) in the 1950s to today's sophisticated image processing and text prediction algorithms, researchers have faced a major challenge: transforming complex visual data into usable textual information. This continuous pursuit has led to an in-depth exploration of various innovative approaches, all aimed at overcoming the inherent challenges of visual documents [7].

Among these approaches, the use of Convolutional Neural Networks (CNNs) has emerged as a particularly promising technique for recognizing patterns in images [8]. CNNs are deep learning models capable of automatically detecting and extracting significant visual features from images, making them highly effective for processing complex visual data, such as digitized documents. Unlike traditional OCR [9], which relies on predefined rules to recognize characters, CNNs can autonomously identify significant patterns and structures, enabling precise recognition of text and other visual elements in documents.

Simultaneously, Natural Language Processing (NLP) techniques offer a complementary approach for analyzing the textual content of documents [10]. NLP methods allow for the understanding and interpretation of human language by extracting meaningful information from raw text. These techniques include tokenization [11], lemmatization [12], named entity recognition [13], and syntactic analysis [14], which facilitate the systematic and effective processing and analysis of text. By combining the visual data-processing capabilities of CNNs with NLP techniques for textual analysis, it is possible to gain a deeper understanding of the content of digitized documents and transform them into actionable information for a variety of applications [15].

In the legal domain, several studies have highlighted the use of machine learning to analyze and classify court decisions, contributing to the automation of tasks that were previously labor-intensive and prone to human error [16][17]. For example, research has shown how clustering and classification techniques can be used to group judicial decisions based on common characteristics, thereby facilitating their analysis and understanding [18]. Additionally, studies have examined the effectiveness of machine learning algorithms such as Support Vector Machines (SVM) and artificial neural networks in predicting legal outcomes [19].

However, most of this research has been conducted on English-language data sets, leaving a gap when it comes to other languages and legal contexts. Future studies could, therefore, focus on adapting and

applying these techniques to specific languages and legal systems, such as Arabic and the Moroccan judiciary, to address this gap in the literature [20]. Furthermore, integrating natural language analysis and NLP techniques could further enhance the ability to understand and interpret judicial decisions across different linguistic contexts [21].

In the area of judicial case type prediction, several studies have explored the use of machine learning techniques to classify decisions, whether in the Moroccan judicial system [22] or in other jurisdictions [23 - 24]. However, it is worth noting that most of this research has focused on analyzing structured data, such as pre-organized databases of judicial decisions.

As such, few studies have specifically addressed the use of machine learning algorithms for prediction based on unstructured data, such as scanned documents [25]. This gap in the literature underscores the importance of our study, which focuses precisely on predicting the types of judicial cases from non-searchable PDF documents in the Arabic language, representing a significant contribution to research in this field.

3.Materials and Methods :

As part of our experiment, we utilized 14,207 Arabic-language judicial decisions from the Administrative Court of Appeal in Marrakech, digitized from paper documents into medium-quality PDF files. These documents, covering 15 types of judicial cases, represent a valuable source of information for our study. Despite the technical challenges inherent to the non-searchable nature of these PDF files, we developed a robust methodology, combining sophisticated data processing tools to analyze these documents and extract relevant information for our research.

3.1.Data structuring :

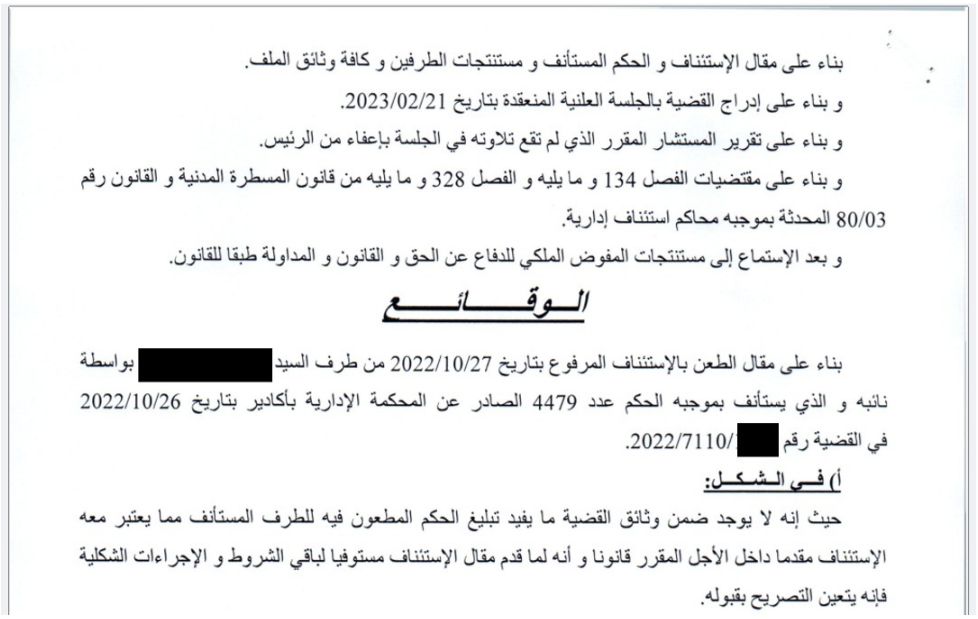
We began by converting each document into a series of images, thus initiating the processing before extracting the text. In our approach, the choice of Python for image processing was crucial to ensure the efficiency and accuracy of our method [26]. Python offers many advantages in the field of artificial intelligence, including its simple syntax, its large developer community, and its wealth of libraries specialized in image processing. After evaluating various available options, we selected the libraries that best met our specific needs.

For converting PDF documents into images, we considered using libraries such as pdfplumber and PyPDF2. However, these alternatives presented limitations in terms of compatibility with certain types of PDF files and error handling during conversion. Instead, we opted for pdf2image to convert each page of our PDF documents into an image and OS to effectively manage system operations and ensure smooth file handling. These libraries were chosen for their robustness and broad compatibility with a variety of PDF formats.

For image preprocessing, we also explored options such as OpenCV and scikit-image, which offer advanced image processing features. However, these libraries were less suited to tasks specific to handling Arabic judicial documents, and their more complex configurations would have increased the complexity of our processing pipeline. Therefore, we chose to use the Python Imaging Library (PIL) [27]. PIL is known for its versatility and ease of use, which allowed us to easily adjust preprocessing parameters such as grayscale conversion, removal of unreadable areas and noise, and sharpening, based on the specific characteristics of Arabic judicial documents.

Regarding text extraction from the images, we evaluated several alternative options such as TesseractOCR and textract. However, these libraries generally offer lower accuracy in recognizing Arabic text compared to pytesseract [28]. Pytesseract is specifically trained to handle non-Latin scripts, including Arabic. Therefore, we favored it due to its proven reputation in Optical Character Recognition (OCR) across various languages, ensuring the accuracy of the data extracted from the judicial decisions. Initially, we saved each extracted text in an HTML file to preserve the source layout and accurate character encoding. Then, we transferred this textual data to an Excel file using the Python libraries xlwt and glob. Each row of this file corresponded to a judicial decision, where column A contained the extracted Arabic text and column B the corresponding case type [29].

To reinforce the applicability of our approach in real-world scenarios, we present below a concrete example of PDF documents before and after processing with the mentioned tools. This example illustrates the effectiveness of artificial intelligence techniques in handling unstructured data, providing a better understanding of the advantages these methods offer in processing judicial documents.



بناء على مقال الإستئناف و الحكم المستأنف و مستنتجات الطرفين و كافة وثائق الملف.

و بناء على إدراج القضية بالجلسة العلنية المنعقدة بتاريخ 2023/02/21.

و بناء على تقرير المستشار المقرر الذي لم تقع تلاوته في الجلسة بإعفاء من الرئيس.

و بناء على مقتضيات الفصل 134 و ما يليه و الفصل 328 و ما يليه من قانون المسطرة المدنية و القانون رقم 80/03 المحدثة بموجبه محاكم استئناف إدارية.

و بعد الإستماع إلى مستنتجات المفوض الملكي للدفاع عن الحق و القانون و المداولة طبقا للقانون.

الـــوقـــــائـــــع

بناء على مقال الطعن بالإستئناف المرفوع بتاريخ 2022/10/27 من طرف السيد [redacted] بواسطة نائبه و الذي يستأنف بموجبه الحكم عدد 4479 الصادر عن المحكمة الإدارية بأكادير بتاريخ 2022/10/26 في القضية رقم [redacted]/2022/7110.

أ) فـــي الـشـكـــل:

حيث إنه لا يوجد ضمن وثائق القضية ما يفيد تبليغ الحكم المطعون فيه للطرف المستأنف مما يعتبر معه الإستئناف مقدما داخل الأجل المقرر قانونا و أنه لما قدم مقال الإستئناف مستوفيا لباقي الشروط و الإجراءات الشكلية فإنه يتعين التصريح بقبوله.

Fig. 1. Data before processing and extraction.

	texte	type
2	بناء على مقال الإستئناف و الحكم المستأنف و مستنتجات الطرفين و كافة وثائق الملف.و بناء على إدراج القضية بالجلسة العلنية المنعقدة بتاريخ 2023/02/21.و بناء على تقرير المستشار المقرر الذي لم تقع تلاوته في الجلسة بإعفاء من الرئيس.و بناء على مقتضيات الفصل 134 و ما يليه و الفصل 328 و ما يليه من قانون المسطرة المدنية و القانون رقم3 المحدثة بموجبه محاكم استئناف إدارية.و بعد الإستماع إلى مستنتجات المفوض الملكي للدفاع عن الحق و القانون و المداولة طبقا للقانون.الوقائع بناء على مقال الطعن بالإستئناف المرفوع بتاريخ 2022/10/27 من طرف السيد [redacted] بواسطةنائبه و الذي يستأنف بموجبه الحكم عدد 4479 الصادر عن المحكمة الإدارية بأكادير بتاريخ 2022/10/26في القضية رقم [redacted]/2022/7110.أ في الشكل:حيث إنه لا يوجد ضمن وثائق القضية ما يفيد تبليغ الحكم المطعون فيه للطرف المستأنف مما يعتبر معهالإستئناف مقدما داخل الأجل المقرر قانونا و أنه لما قدم مقال الإستئناف مستوفيا لباقي الشروط و الإجراءات الشكليةفإنه يتعين التصريح بقبوله.	قضاء الإلغاء

Fig. 2. Data after processing and extraction.

Note: Proper names have been masked in the screenshots to ensure the confidentiality of sensitive personal data.

This final Excel file constituted our dataset, which we used for machine learning to predict the different types of judicial cases. This methodology allowed us to effectively structure the extracted data for subsequent analysis while adopting a standardized format to train our machine learning models.

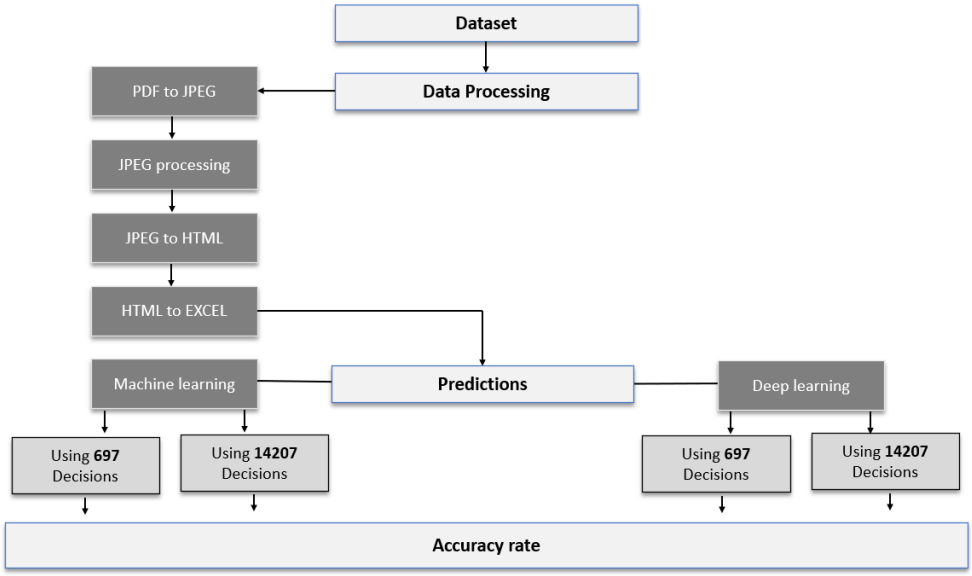


Fig. 3. Proposed Model Schema

Our objective was clear from the start : First, to evaluate the performance of machine learning and then explore deep learning. This comparative approach, initially conducted on 697 decisions and later on 14207 decisions, allowed us to identify the most suitable method for our study.

3.2.Machine Learning Phase :

In the first phase of our study, focusing on machine learning, we implemented three classification algorithms [30] to predict the case type for each judicial decision. Machine learning, a subset of artificial intelligence, enables computers to learn from data and improve their performance without being explicitly programmed. Its classification algorithms, such as decision trees, Naive Bayes, and SVM, are used to categorize data based on their features.

- Decision Tree : We began our analysis with the decision tree, a widely used method in data analysis and machine learning. This algorithm segments the data into subgroups and establishes a series of decision rules based on the characteristics of each group.
- Naive Bayes : We also employed the Naive Bayes model, a statistical classification method based on probability and Bayes' theorem, which assumes the independence of all features.

Support Vector Machine (SVM) : Finally, we utilized the SVM method, a popular classification approach that involves finding a hyperplane to separate the data into different classes.

3.3.Deep Learning Phase :

The second phase of our study was dedicated to deep learning [31], a sub-discipline of machine learning that relies on the use of artificial neural networks to analyze and interpret complex data. In contrast to traditional machine learning methods, deep learning can automatically extract relevant features from raw data. This capability makes it especially effective for tasks such as image recognition, natural language processing, and more.

We focused on two specific algorithms :

- Multi-Layer Perceptron (MLP) : MLP is an artificial neural network model consisting of multiple layers of nodes, including an input layer, one or more hidden layers, and an output layer. MLP can learn complex relationships between the data by adjusting the weights of the connections between the layers during the training phase.
- Probabilistic Neural Network (PNN) : Unlike MLP, which uses a supervised learning approach, PNN is a neural network model based on a probabilistic approach. It is mainly used for classification and operates by estimating the posterior probabilities of classes for each input instance based on the provided training data.

3.4.Model Evaluation :

After applying these various algorithms to predict the type of each judicial case, we evaluated the accuracy of our predictions using a sophisticated method called cross-validation. This complex technique involved several key steps : First, our data were divided into distinct training and validation sets. Then, we trained our models on the training set, allowing them to learn the patterns and features inherent in our data. Finally, we tested the performance of our models on the validation set, which provided an unbiased assessment of their accuracy. Through this rigorous methodology, we not only estimated the reliability of our predictions but also identified the best-performing classification algorithm for our specific dataset.

4.Results and Discussion :

To evaluate the text extraction method, a sample of 10 judicial decisions was manually reviewed to verify that the extracted words were correct and without omissions. The words were then compared to an online Arabic dictionary (arabdict.com) to ensure their validity. The correct extraction rate was calculated and revealed an average accuracy of 95%, with errors mainly related to segmentation issues, such as words being stuck together or missing spaces. This result confirms the effectiveness of the extraction method applied to digitized judicial documents.

With the successful text extraction process established, we proceeded to apply machine learning and deep learning techniques to predict the types of judicial cases based on the extracted data.

4.1.Using machine Learning :

The evaluation of the system's ability to identify the types of legal cases from the decisions using machine learning revealed a significant variation in accuracy across different classification algorithms.

In the experiment on 697 decisions, SVM stood out with the best performance, achieving an accuracy rate of 91%, followed by the decision tree at 81%, while Naïve Bayes recorded a much lower score of only 30%. In the experiment on 14207 decisions, SVM again produced the best results with an accuracy of 97%, followed by the decision tree at 92%, while Naïve Bayes, though improved, still showed inferior performance with a rate of 88%. This performance disparity highlights the distinct characteristics of each algorithm and their ability to handle the data.

Naive Bayes, despite its assumption of conditional independence between features, can be effective in certain situations. Its decision function can be formulated as follows:

$$f_{NB}(x) = \text{argmax}_y \, P(Y = y) \prod_{i=1}^{n} P(X_i = x_i | Y = y) \quad (1)$$

Here, *P(Y = y)* represents the prior probability of class **Y**, while *P(Xi = xi | Y = y)* is the conditional probability of observing feature **Xi** given class **y**.

Although this method relies on a conditional independence assumption, which may seem unrealistic in real-world contexts where features can be interdependent, Naive Bayes has the advantage of being fast to train and can deliver satisfactory results when this assumption is reasonably met.

However, our experiment showed that Naive Bayes achieved an accuracy of 30% when analyzing 697 decisions, but reached 88% with the analysis of 14207 decisions. This performance difference can be explained by the fact that with a larger number of decisions, the model benefits from a greater volume of data, which can improve the accuracy of conditional probability estimates and mitigate the impact of the conditional independence assumption. In other words, Naive Bayes can offset its simplifying assumptions when the dataset is sufficiently large to capture a wider range of variations and dependencies between features.

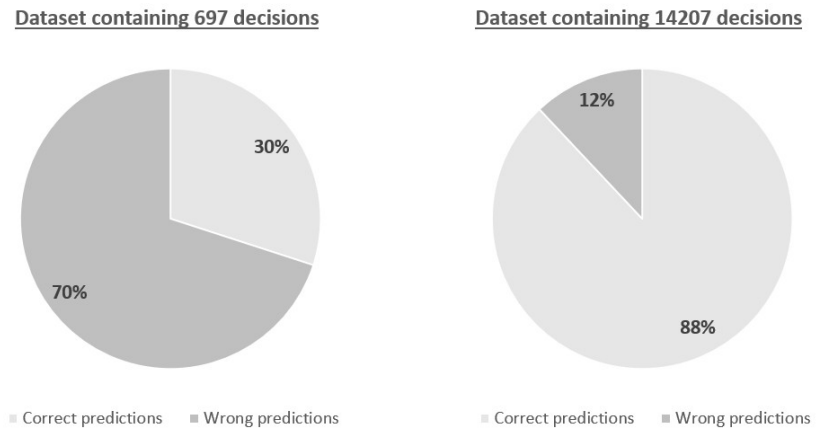


Fig. 4. Graph of Predictions Using the "Naive Bayes" Algorithm

Conversely, the decision tree, although relatively simple to interpret, can be prone to overfitting, limiting its ability to generalize to new data. Despite this, its approach of dividing the feature space into segments defined by simple rules makes it suitable when model transparency is crucial and facilitates the identification of decisions made by the algorithm.
The functioning of a decision tree can be formulated as a series of decision rules based on the characteristics of each group. Mathematically, this can be represented as follows :

$$f_{DT}(x) = \begin{cases} 1 & \text{if condition} \\ 0 & \text{else} \end{cases} \quad (2)$$

Where **x** represents the features of the instance to be classified, and the conditions are defined by the divisions of the tree.

In the experiment with 697 decisions, the decision tree achieved an accuracy rate of 81%. However, in the analysis of 14207 decisions, its accuracy increased to 92%. This improvement can be explained by the fact that a larger volume of data allows the decision tree to better learn and adjust its segmentation rules, thereby reducing overfitting and improving generalization to new data. With a larger dataset, the decision tree can capture more variations and refine its classification criteria, leading to enhanced performance.

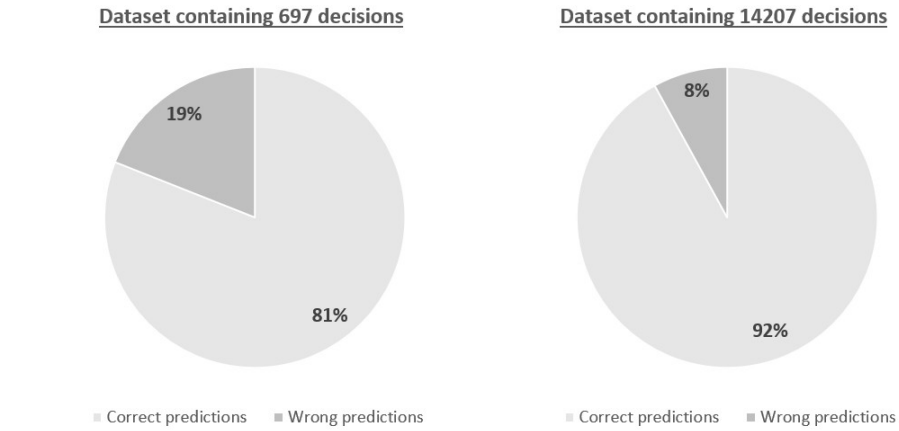


Fig. 5. Graph of Predictions Using the "Decision Tree" Algorithm

The SVM, on the other hand, stands out for its ability to better generalize on nonlinear data thanks to the use of appropriate kernels. It is also robust against overfitting. However, its computational complexity can be a drawback, especially for large datasets where the computation time can be significant.

In the case of SVM, the decision function can be represented as follows :

$$f_{SVM}(x) = \text{sign}\left(\sum_{i=1}^{n} \alpha_i y_i K(x_i, x) + b\right) \quad (3)$$

Here, **αi** are the Lagrange coefficients, **yi** are the class labels, *K(xi,x)* is the kernel that measures the similarity between the support vectors **xi** and the input vector **x**, and **b** is the bias. This equation shows that the classification decision depends on the weighted sum of the dot products between the support vectors xi and the input vector **x**, plus a bias term **b**.

In the analysis of 697 decisions, the SVM achieved an accuracy rate of 91%. However, in the analysis of 14207 decisions, its accuracy increased to 97%. This improvement can be attributed to the fact that a larger volume of data allows the SVM to better capture the nuances and complexities of the data, enabling a more precise separation of the classes. With a larger dataset, the SVM can better adjust its parameters and refine the decision boundaries, significantly enhancing its performance.

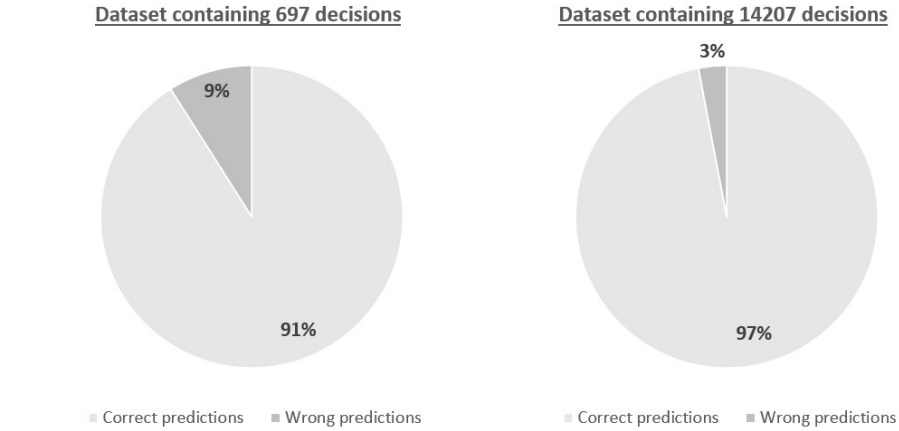


Fig. 6. Graph of Predictions Using the "SVM" Algorithm

4.2.Using deep Learning :

In the second phase of our study, we evaluated two deep learning algorithms: the Multi-Layer Perceptron (MLP) and the Probabilistic Neural Network (PNN) on our dataset of judicial decisions. The results were impressive, with a prediction accuracy of 100% for both algorithms in the experiment on 697 decisions, but an accuracy of 96% in the experiment on 14207 decisions. This variation in performance can be attributed to the specifics of each algorithm.

The MLP, with its hidden layers allowing it to learn hierarchical representations of the data, is capable of capturing complex relationships between the features of judicial decisions. Its decision function can be mathematically represented as follows :

$$f_{MLP}(x) = \arg\max_{y} \sigma\left(\sum_{i=1}^{N} W_i^{(2)} \cdot \sigma\left(\sum_{j=1}^{M} W_{ij}^{(1)} \cdot x_j + b_i^{(1)}\right) + b^{(2)}\right) \quad (4)$$

Where **x** represents the features of the instance to be classified, *Wij(1)* and *Wi(2)* are the weights of the connections between the hidden layers and the output layer, *b(1)* and *b(2)* are the biases, and **σ** is the activation function.

In the experiment with 697 decisions, the MLP achieved a prediction accuracy of 100%. However, in the analysis of 14207 decisions, its accuracy slightly decreased to 96%. This drop in performance may be due to the increased complexity of the data in the larger dataset, which can make the model more susceptible to overfitting or require finer optimization of parameters. With a larger dataset, the MLP has to handle greater variability and complexity, which can affect its ability to generalize as well as with a smaller dataset.

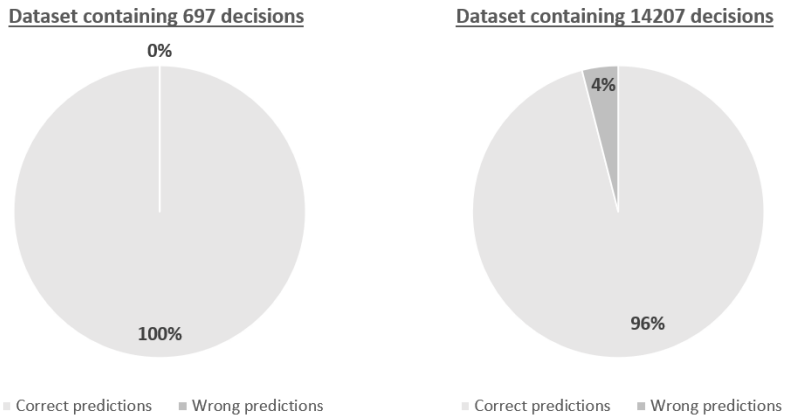


Fig. 7. Graph of Predictions Using the "MLP" Algorithm

The PNN, on the other hand, is based on computing the conditional probabilities of the classes for each input instance. This approach allows it to estimate the posterior probabilities of the classes, which is particularly useful in contexts where a probabilistic assessment is crucial.

Its decision function can be formulated as follows :

$$f_{PNN}(x) = \arg\max_y P(Y = y) \prod_{i=1}^{N} P(X_i = x_i | Y = y) \quad (5)$$

Where *P(Y=y)* represents the prior probability of class **y**, and *P(Xi=xi /Y=y)* is the conditional probability of observing the feature **xi** given class **y**.

During the analysis of 697 decisions, the PNN achieved a prediction accuracy of 100%. However, in the analysis of 14207 decisions, its accuracy decreased to 96%. This decline can be attributed to the increased complexity of the larger dataset, which may make the model more sensitive to variations and noise in the data. With a larger volume of data, the PNN may struggle more to accurately estimate the conditional probabilities due to the increased diversity of instances, which can reduce its overall performance.

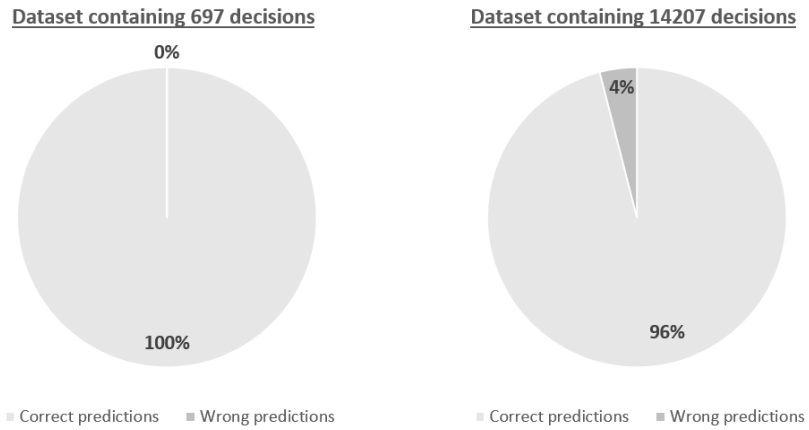


Fig. 8. Graph of Predictions Using the "PNN" Algorithm

The generalization capability and accuracy of these algorithms are supported by their respective abilities to learn complex patterns from training data. The MLP can learn hierarchical representations, while the PNN uses a probabilistic approach to estimate the posterior probabilities of classes for each input instance.

4.3. Summary of key findings and comparative analysis :

The results revealed notable differences between machine learning and deep learning algorithms, both for the analysis of 697 decisions and for the analysis of 14207 decisions.

For machine learning algorithms :

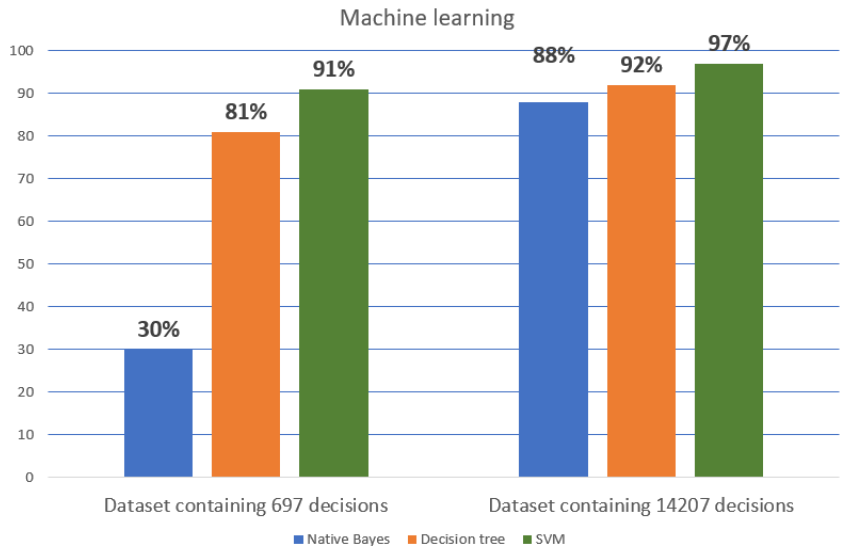


Fig. 9. Graph comparing different machine learning algorithms used in our predictions

For deep learning algorithms :

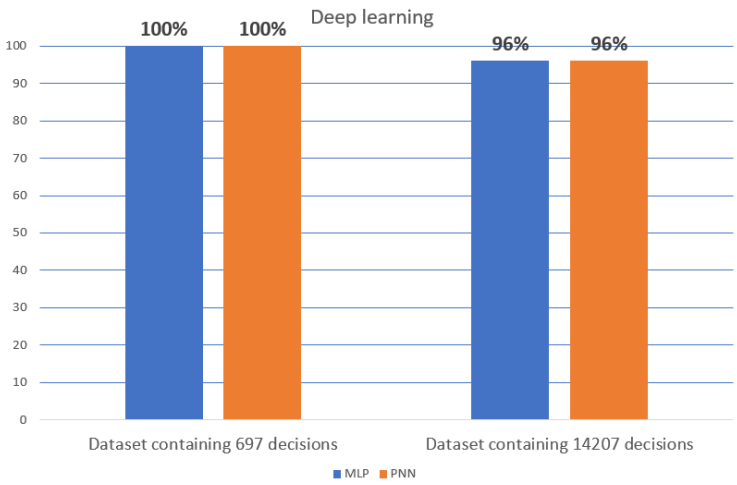


Fig. 10. Graph comparing different machine learning algorithms used in our predictions

The results indicate that machine learning algorithms, such as SVM, Decision Tree, and Naive Bayes, improved their accuracy with a larger dataset. This improvement is due to the enhanced ability of these algorithms to detect more complex patterns and reduce errors when provided with a greater volume of data. In particular, SVM demonstrated a remarkable accuracy of 97% with the larger dataset, while Decision Tree and Naive Bayes also showed significant improvements compared to their performance with the smaller dataset. These results highlight the importance of having a sufficient amount of data to optimize the accuracy of machine learning models.

Deep learning algorithms, notably MLP and PNN, maintained high performance across both datasets. Their ability to achieve perfect accuracy of 100% on 697 decisions, as well as 96% accuracy on 14,207 decisions, suggests they are particularly effective at capturing the nuances and complex structures in the data. However, the perfect accuracy of MLP and PNN on the smaller dataset may indicate a potential risk of overfitting, although their performance remains strong with the larger dataset.

This risk of overfitting occurs when the model memorizes the training data instead of learning underlying patterns, which compromises its ability to generalize to new data. This is particularly concerning with a small dataset of 697 decisions, as the model might have learned specific details unique to these data, limiting its ability to predict accurately on larger or new datasets.

To mitigate this risk, several measures were implemented:

- Cross-validation: This method ensures that the model generalizes well across different subsets of the data.
- Regularization: Techniques such as dropout were used to prevent the model from becoming too specialized to the training set.
- Data augmentation: Where possible, variations were introduced in the documents to diversify the training set and improve generalization.
- Evaluation on unseen data: Model performance was measured on a separate test set to ensure that the model was not overfitting.

Despite these precautions, a perfect accuracy of 100% on a small dataset should be interpreted with caution. This is why additional tests on larger datasets were necessary to confirm the robustness of the models.

This experience shows that both machine learning and deep learning algorithms can produce excellent results when used with sufficiently large datasets. Deep learning algorithms, in particular, are highly effective at understanding complex relationships in the data, especially with larger datasets. However, attention must be paid to the risk of overfitting, especially with smaller datasets. These findings emphasize the importance of selecting the right algorithms based on the characteristics of the data to achieve accurate and reliable predictions.

4.4.Impact of the experiment on a real case (The Administrative Court of Appeal in Marrakech) :

The experiment conducted on 14,207 decisions from the Administrative Court of Appeal in Marrakech, classified into 15 distinct types, had a direct and measurable impact on the court's operations, proving its effectiveness in practice. The experience modernized and optimized several aspects of judicial management, notably by speeding up case processing and facilitating access to similar decisions, thereby enhancing the efficiency and consistency of the court.

Among these improvements, the following aspects stand out:

- Digitization of Decisions and Access to a Searchable Database : The digitization of decisions transformed how judges and clerks access information. Previously, decisions were stored in paper format or non-searchable digital files, making the search for specific decisions time-consuming and imprecise. Thanks to this digitization, a searchable database was created, allowing magistrates to quickly retrieve decisions based on specific criteria, such as keywords or case types. This greatly facilitates their work, enabling them to save time in analysis and make more informed decisions.
- Quick Identification of Similar Cases : Access to this database allows judges to rapidly identify similar decisions, which is crucial

in legal systems based on jurisprudence. For instance, when a judge is handling a case, they can quickly search for previous decisions with similarities to guide their own ruling. This alignment with established precedents fosters greater consistency and fairness in the verdicts delivered.

- Anonymization of Decisions for Public Sharing : A key benefit of this digitization is the ability to easily anonymize decisions before sharing them with the public. Anonymization involves removing or masking personal information (names, addresses, etc.) to protect the privacy of the individuals involved, in accordance with data protection regulations. This allows certain decisions to be made accessible to the public or researchers without compromising the confidentiality of the parties involved, thus enhancing the transparency of the judicial system by enabling controlled access to decisions.
- Classification of Older Decisions : Digitization is not limited to new decisions; it has also enabled the classification of older decisions that were not clearly identified or categorized. For example, decisions made in the past without precise classification were organized based on their type or subject. This makes the management of judicial archives more efficient and allows judges to easily consult older decisions that were previously difficult to access or unclassified.

Efficiency, Speed, and Reduction of Human Errors : The digitization of decisions has significantly accelerated and optimized the entire process. Whereas judges and clerks previously had to manually search through documents or files, automated searches in the database now allow for access to relevant information in a matter of seconds. By eliminating manual handling of documents, this greatly reduces the risk of human errors, whether in data entry, omissions, or classification mistakes.

In summary, this initiative has brought considerable advantages in terms of time savings, accuracy, and security, while contributing to the consistency of judicial decisions and the overall improvement of case management.

5. Ethical considerations and confidentiality :

In the context of this study conducted at the Administrative Court of Appeal in Marrakech, the judicial documents used were not anonymized. This is due to the fact that the personal information contained within them is essential for internal research carried out by judges and court personnel. To ensure the effectiveness of these searches, judicial decisions must remain accessible to authorized individuals, particularly for queries based on criteria such as name, date, or other pertinent information.

To guarantee the confidentiality and security of the data, access to the digitized documents is strictly limited to court personnel. No personal data or sensitive information leaves the premises of the Court of Appeal, and these files are not accessible to the public. All necessary precautions have been taken to ensure that this information remains protected from unauthorized disclosure.

For decisions intended for future public use, particularly for researchers and law students, a process of anonymizing the digitized decisions has been implemented. This process involves removing or masking all personal information, such as names, addresses, and other identifying data, to protect the privacy of the individuals involved. The goal is to make these decisions accessible to a broader audience while strictly adhering to data protection regulations. Once anonymized, these decisions can be shared safely, ensuring both transparency and confidentiality.

While artificial intelligence is used in this context to facilitate internal research and analysis, we have adhered to the data protection regulations applicable within the court. This approach aims to enhance the efficiency of judicial processes without compromising the confidentiality of sensitive information.

6.Conclusion :

In conclusion, this doctoral research explored the application of machine learning to the analysis of Arabic-language judicial decisions from non-searchable PDF documents. One of the main challenges addressed in this study was the extraction of Arabic text, an essential step that was successfully completed, enabling the models to be trained effectively.

The results showed a significant difference in the performance of machine learning and deep learning algorithms depending on the datasets used. For the first set of 697 decisions, SVM achieved an accuracy of 91%, while MLP and PNN reached a perfect accuracy of 100%. However, this perfect accuracy may indicate a risk of overfitting, where the models memorize the data rather than generalize effectively. This risk was less pronounced for the larger set of 14,207 decisions, where SVM achieved 97% accuracy, and deep learning algorithms reached 96%, reflecting better generalization.

This research represents a major breakthrough in processing non-searchable Arabic documents, demonstrating the effectiveness of machine learning for analyzing judicial decisions. It also opens up new perspectives, particularly through the integration of online learning techniques that could continuously update models to keep up with the evolving legal language and judicial practices, ensuring consistently accurate and up-to-date analyses. The developed models can be easily transferred to other administrative courts or jurisdictions, whether in

Morocco or internationally, simplifying the automation of processing and classification of decisions.

Finally, integrating these algorithms into judicial systems could create predictive analysis tools capable of identifying trends in similar cases, thus providing valuable support to judges in decision-making.

Rreferences :

1. C.M. Dahl, T.S.D. Johansen, E.N. Sørensen, C.E. Westermann, S.F. Wittrock, *Appl. Mach. Learn. Digit. Document* **14**, 7 (2021)
2. S.V. Nguyen, D.A. Nguyen, L.S.Q. Pham, *Digit. Adm. Docs Pract.* **21**, 4 (2021)
3. X. Ding, D. Wen, L. Peng, C. Liu, *Doc. Digit. Technol. Appl. Digit. Lib. China* **8**, 15 (2004)
4. M. Charfi, W. Boussella, M.A. Alimi, *Intell. Syst. Digit. Hist. Arab. Docs* **5**, 9 (2007)
5. E. Mohammed, E. Mustapha, M. Azhari, *Mach. Learn. Predict. Publ. Prosecut. Judges Decis. Moroccan Courts* **13**, 27 (2023)
6. S. Mukherjee, H. Tyagi, P. Tyagi, N. Singh, S. Bhardwaj, *OCR Python Appl.* **18**, 23 (2023)
7. E. Gunaydın, B. Gencturk, C. Ergen, M. Köklü, *Digit. Arch. Invoices Deep Learn. Text Recognit.* **20**, 11 (2020)
8. K.C. Nguyen, C.T. Nguyen, *Doc. Digit. Deep Conv. Neural Netw.* **21**, 12 (2020)
9. S. Stoliński, W. Bieniecki, *Appl. OCR Syst. Proc. Digit. Paper Docs* **10**, 3 (2011)
10. A.Farghaly, K. Shaalan, *Arab. Nat. Lang. Process. Challenges Solut.* **7**, 14 (2009)
11. S.J. Mielke, Z. Alyafeai, E. Salesky, C. Raffel, M. Dey, M. Gallé, A. Raja, C. Si, W.Y. Lee, B. Sagot, S. Tan, *Hist. Open-Vocab. Model. Tokeniz. NLP* **32**, 6 (2021)
12. D. Khyani, *Interpret. Lemmatiz. Stem. Nat. Lang. Process.* **25**, 19 (2021)
13. N. Alsaaran, M. Alrabiah, *Arab. Named Entity Recognit. BERT-BGRU Approach* **30**, 21 (2021)
14. Y.T. Zhou, *Nat. Lang. Process. Improv. Deep Learn. Neural Netw.* **16**, 24 (2022)
15. N.I. Widiastuti, *Conv. Neural Netw. Text Min. Nat. Lang. Process.* **22**, 8 (2019)
16. F. Wei, *Empir. Study Deep Learn. Text Classif. Leg. Doc. Rev.* **9**, 27 (2018)
17. T.T. Cheng, *Inf. Extract. Leg. Docs* **6**, 11 (2009)
18. R.S. de Oliveira, E.G.S. Nascimento, *Brazil. Court Docs Cluster. NLP Transformers* **23**, 7 (2022)

19. R. Sil, A. Roy, *Novel Argument Legal Predict. Mach. Learn.* **13**, 12 (2020)
20. E. Schweighofer, A. Rauber, M. Dittenbach, *Autom. Text Represent. Classif. Labeling Eur. Law* **30**, 5 (2001)
21. K. Faidi, R. Ayed, I. Bounhas, B. Elayeb, *Compar. Arab. NLP Tools Hadith Classif.* **18**, 14 (2015)
22. A.I. Yahia, *Arab. Text Classif. Legal Domain* **21**, 17 (2019)
23. H. Chen, *Compar. Study Autom. Leg. Text Classif. Random Forest Deep Learn.* **33**, 9 (2022)
24. N. Limsopatham, *Leveraging BERT Leg. Doc. Classif.* **20**, 6 (2021)
25. D. Zhang, C. Yin, J. Zeng, X. Yuan, P. Zhang, *Struct. Unstruct. Data Predict. Models Deep Learn.* **40**, 16 (2021)
26. R. Chityala, S. Pudipeddi, *Image Process. Acquis. Python* **27**, 11 (2020)
27. R.R. Asaad, R.I. Ali, Z.A. Ali, A.A. Shaaban, *Image Process. Python Libraries* **29**, 13 (2023)
28. J.M. Jayoma, E.S. Moyon, E.M.O. Morales, *OCR Based Doc. Archiv. Indexing PyTesseract* **18**, 22 (2020)
29. T.F. Gharib, *Arab. Text Classif. SVM* **11**, 7 (2009)
30. M. Ikonomakis, *Text Classif. Mach. Learn. Tech.* **23**, 5 (2005)
31. S. Li, B. Gong, *Word Embedd. Text Classif. Deep Learn.* **39**, 14 (2021)