Online-Journals.org

iJOE International Journal of Online and Biomedical Engineering

iJOE | eISSN: 2626-8493 | Vol. 21 No. 2 (2025) | OPEN ACCESS

https://doi.org/10.3991/ijoe.v21i02.52639

PAPER

Prediction of Medical Pathologies: A Systematic Review and Proposed Approach

Chaimae Taoussi[1](✉), **Imad Hafidi**[1], **Abdelmoutalib Metrane**[2]

[1]University Sultan Moulay Slimane, Beni-Mellal, Morocco

[2]Cadi Ayyad University, Marrakech, Morocco

chaimae.taoussi@usms.ac.ma

ABSTRACT

Healthcare is essential in every society, and the adoption of innovative technologies such as artificial intelligence (AI), big data, machine learning (ML), and deep learning (DL) is revolutionizing medical practices by enabling innovative approaches to pathology prediction and clinical decision-making. This systematic review examines 61 key articles published between 2018 and 2024 to evaluate the state of the art in medical data processing and pathology prediction. Based on this review, we identify critical challenges in current methodologies, including data integration and interpretability. To address these issues, we propose an integrated framework combining data collection, pre-processing, mapping, and clustering with advanced analytics. This approach aims to streamline the medical data pipeline, enhance diagnostic processes, and provide a foundation for future research and clinical implementation.

KEYWORDS

artificial intelligence (AI), big data, healthcare, data mining, natural language processing (NLP), predictive models

1 INTRODUCTION

The term "Big Data" refers to the computational capacity required to manage the vast and complex datasets generated from various sources, including structured, semi-structured, and unstructured data. These systems increasingly rely on automation driven by artificial intelligence (AI), which has revolutionized decision-making processes, particularly in diagnostics [1]. In healthcare, leveraging big data from sources such as medical records, patient files, and examination results has enabled significant advancements for both healthcare practitioners and patients [2]. Electronic health records (EHR), in particular, hold immense potential for transforming biomedical research, offering data critical to advancing precision medicine and improving medical treatments [3].

Despite these advancements, significant challenges persist in fully leveraging the potential of healthcare big data. Data fragmentation from multiple sources and the

Taoussi, C., Hafidi, I., Metrane, A. (2025). Prediction of Medical Pathologies: A Systematic Review and Proposed Approach. *International Journal of Online and Biomedical Engineering (iJOE)*, 21(2), pp. 121–136. https://doi.org/10.3991/ijoe.v21i02.52639

Article submitted 2024-10-02. Revision uploaded 2024-12-06. Final acceptance 2024-12-07.

diversity of formats (structured, semi-structured, and unstructured) complicate their analysis and integration into clinical systems [1], [2]. For instance, while EHRs alone represent a valuable resource, substantial efforts are required to harmonize them with other types of clinical data [3]. Furthermore, the absence of standardized data structures hinders the extraction of meaningful insights, making clinical decision-making inefficient and labor-intensive [4]. These limitations are compounded by the sheer volume of healthcare data, which demands sophisticated systems capable of processing complex datasets efficiently while ensuring accessibility [6].

In today's rapidly evolving landscape, healthcare professionals also require digital literacy skills to operate technological tools and provide tech-enabled services [5]. A lack of expertise in this area can compromise patient safety and increase the risk of errors [6]. Furthermore, data mining techniques have proven invaluable for extracting insights from clinical databases, offering decision support to predict various conditions with high accuracy. These techniques are particularly useful in designing clinical support systems capable of detecting hidden patterns and relationships within medical data [7].

Current methodologies, although promising, exhibit critical limitations. Interoperability gaps prevent seamless communication between heterogeneous data formats, obstructing the efficient exchange of information [8]. Moreover, many machine learning (ML) algorithms operate as "black boxes," lacking interpretability, which undermines their adoption by healthcare professionals who require trust and transparency in clinical tools [9]. Additionally, existing solutions are often tailored to specific diseases, limiting their generalization and scalability across diverse healthcare scenarios [10]. These limitations underscore the need for frameworks that not only address data integration but also provide interpretability and broad applicability.

To address these challenges, this study presents a systematic review of recent advancements in AI, ML, and big data applied to medical pathology prediction. Based on this analysis, an integrated approach is proposed, combining data collection, pre-processing, mapping, and classification to improve diagnostic efficiency and facilitate adoption in clinical environments [11].

2 METHODOLOGY

This systematic review adheres to the preferred reporting items for systematic reviews and meta-analyses (PRISMA) guidelines, as illustrated in Figure 1, to ensure transparency, reproducibility, and comprehensiveness. These guidelines provide a standardized framework, which was applied to all stages of the review process, including data collection, screening, inclusion, and exclusion criteria.

2.1 Study identification and search strategy

To identify relevant studies, an extensive literature search was conducted across major databases, including Scopus, PubMed, and Google Scholar. The search focused on articles published between 2018 and 2024 and used combinations of keywords such as "Data Collection," "Data Preprocessing," "Mapping Medical Data," and "Classification and Clustering" alongside terms such as "Artificial Intelligence" and "Medical Pathology Prediction." Studies were selected based on their relevance to predefined research questions (RQ) and their contribution to addressing challenges in health informatics.

2.2 Research questions and inclusion/exclusion criteria

The review addressed five primary RQs, summarized in Table 1, focusing on data collection, preprocessing, mapping techniques, classification methods, and AI applications in predicting medical pathologies. Inclusion criteria targeted studies that focused on data collection, preprocessing, mapping, clustering, or AI applications in health informatics. Exclusion criteria eliminated literature reviews, non-full-text articles, or studies irrelevant to the research questions.

Table 1. Specific research questions

ID	Research Questions
RQ 1	What are the data collection methods applied in medical informatics?
RQ 2	What are the data preprocessing methods applied in medical informatics?
RQ 3	What are the data mapping techniques applied in medical informatics?
RQ 4	What is the classification and clustering methods applied in medical informatics?
RQ 5	How can artificial intelligence be used to predict medical pathologies?

2.3 Results

A total of 746 papers, published between 2018 and 2024, were retrieved during the search process. After the initial screening based on titles and abstracts, 562 articles were deemed irrelevant and excluded. Subsequently, 184 full-text articles were assessed for eligibility based on the inclusion and exclusion criteria. After this review, 123 articles were excluded for reasons such as irrelevance to the RQs, lack of focus on the key areas of our study, duplicates, or methodological issues. After applying inclusion and exclusion criteria, 61 studies were ultimately included for this systematic review. Figure 1 provides a visual representation of the search and selection process.

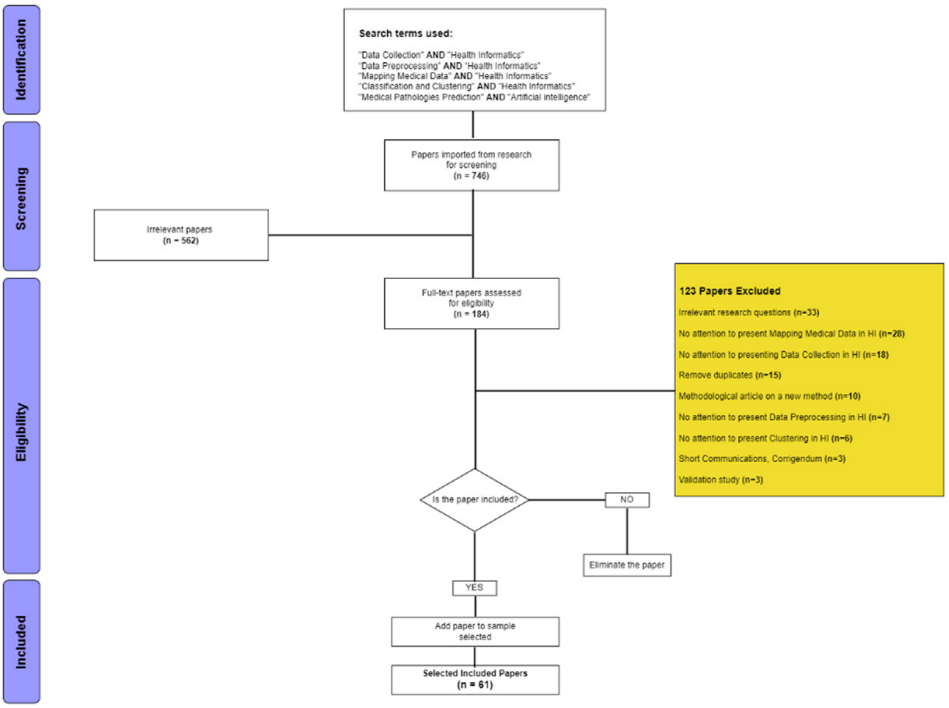

Fig. 1. Search methodology followed in the study

2.4 Report review

RQ1: What are the data collection methods applied in medical informatics? Data collection plays a pivotal role in enhancing healthcare outcomes, including prevention, diagnosis, and treatment. Seol et al. [8] developed an NLP solution for extracting medical events from EHRs, effectively managing both text and metadata using XML files. The use of big data in healthcare enables the identification of patterns, transforming large datasets into actionable insights for decision-making [9]. Hariri et al. [9] emphasized the importance of integrating data from multiple sources to streamline disease prediction and prevention efforts. Similarly, Munoz-Gama et al. [10] highlighted the value of EHR systems in capturing essential healthcare process data, with event logs serving as a key resource for process mining. These approaches demonstrate how well-designed data collection methods can optimize healthcare analytics.

RQ2: What are the data preprocessing methods applied in medical informatics? Preprocessing is a critical step in ensuring data quality and accuracy by performing tasks such as cleaning, integration, and transformation. Modi et al. [11] investigated techniques such as named entity recognition (NER) and relation extraction (RE) to identify important medical concepts, thereby enhancing the accuracy of predictive models. The natural language toolkit (NLTK) is frequently used for text cleaning and processing [12]. For instance, Jha et al. [13] used it for text-to-emoticon conversion, while Yao et al. [14] applied NLTK to estimate emotional scores from sentences [15].

ASSALE et al. [16] emphasized the significance of natural language processing (NLP) in managing unstructured data within EHRs. Similarly, Carchiolo et al. [17] showcased the use of AI for analyzing digitized medical prescriptions. Despite its advantages, data mining for complex datasets such as EHRs still faces challenges related to high-dimensional data, as noted by Hariri et al. [18]. Nevertheless, predictive data mining remains crucial for developing models that help physicians optimize diagnostic and treatment strategies [19].

The general architecture for text engineering (GATE) is a widely adopted framework for processing medical data, particularly for information extraction from English documents, although it also supports other languages through shared user-built dictionaries [20]. Amjad et al. [21] leveraged GATE for multilingual sentiment analysis, demonstrating its utility across diverse linguistic contexts. Pezoulas et al. [22] developed an automated data curation system to improve the quality of medical datasets, while Goldberg et al. [23] illustrated the integration of NLP with ML to forecast critical aspects of psychotherapy. These methods collectively highlight the vital role of pre-processing in addressing the challenges posed by high-dimensional medical datasets and advancing healthcare informatics.

RQ3: What are the data mapping techniques applied in medical informatics? Data mapping techniques facilitate the standardization and integration of biomedical data, enabling more effective healthcare analytics. Topaz et al. [24] developed NimbleMiner, a clinical text mining system that leverages NLP to map biomedical terminology. Al-Hroob et al. [25] proposed an approach for automatically identifying actors and actions in natural language systems, while Wang et al. [26] reviewed clinical information extraction applications that encode textual data into structured formats.

The unified medical language system (UMLS) is a key tool for biomedical data standardization. Kim et al. [27] conducted a bibliometric analysis of UMLS-related publications, and Gorrell et al. [28] developed Bio-YODIE, a system designed to annotate documents using UMLS concepts. Similarly, Abbas et al. [29] proposed an algorithm that integrates UMLS Terminology Services for extracting concepts from clinical discharge summaries.

These tools have been successfully applied across domains such as psychology [30] and oncology [31], highlighting their versatility in medical data mapping.

RQ4: What are the classification and clustering methods applied in medical informatics? In healthcare, clustering is extensively used to identify subgroups of patients with similar profiles, categorize diverse patient populations based on their diagnostic histories, and uncover phenotypic clusters along with their associated risk factors. Maurits et al. [32] developed a framework utilizing longitudinal EHRs for patient stratification, while Ricciardi et al. [33] employed Random Forest algorithms to classify stages of Parkinson's disease. Wang et al. [34] utilized latent dirichlet allocation (LDA) to detect latent disease clusters within EHR data.

Kadhim et al. [35] demonstrated enhanced classification accuracy through a preprocessing system incorporating TF-IDF and cosine similarity techniques. Kashina et al. [36] validated the effectiveness of logistic regression for medical text classification tasks. Meanwhile, Jerlin et al. [37] optimized disease classification by integrating the multiple kernel support vector machine (MKSVM) with the fruit fly optimization algorithm (FFOA). Huang et al. [38] further advanced clustering with a community-based federated learning (CBFL) algorithm aimed at improving learning efficiency on electronic medical record datasets.

Deep learning (DL) techniques are also widely employed for classification and clustering. Desai et al. [39] applied DL to classify biomedical data, while Wood et al. [40] integrated homomorphic encryption to ensure data security within Naive Bayes (NB) classification models. Additionally, Naegelin et al. [41] highlighted the effectiveness of gradient boosting models in classifying stress levels based on mouse and keyboard usage patterns.

To address the complexity of medical datasets, optimization techniques such as feature selection and dimensionality reduction have proven invaluable. For instance, principal component analysis (PCA) and grid search focus models on the most critical features, reducing computational demands and improving predictive accuracy. These strategies are particularly effective in enhancing diabetes prediction models and overall AI applications in healthcare [42].

RQ5: How can artificial intelligence be used to predict medical pathologies? Artificial intelligence plays a transformative role in precision medicine by integrating multimodal and multi-omics data to enable patient-specific decision-making [43]. DL models, particularly convolutional neural networks (CNNs), are essential in predicting and classifying outcomes for both individuals and larger populations. CNNs have demonstrated excellent performance in diagnosing neurodegenerative conditions such as Alzheimer's disease and classifying disease stages from MRI scans, providing healthcare professionals with accurate tools for early diagnosis [45]. Similarly, ML techniques such as XGBoost have achieved up to 98% accuracy in oncology diagnoses, underscoring their impact on early disease detection and clinical decision-making [44].

Ensemble learning methods, including stacking regression, are highly effective in managing chronic diseases. For example, these methods have improved the accuracy of diabetic nephropathy predictions by combining multiple models, enabling early-stage detection and intervention [46]. CNN models have also outperformed traditional techniques such as k-nearest neighbors (KNN) in disease prediction accuracy [47]. Logistic regression remains a competitive approach, with studies by Shipe et al. [48] and Nusinovici et al. [49] demonstrating its comparable effectiveness to advanced ML models for forecasting chronic disease risks.

Deep learning applications extend beyond diagnostics to mortality predictions. Ramzan et al. [50] applied DL techniques to predict Alzheimer's disease using MRI data, while Ye et al. [51] utilized AI to estimate mortality rates in diabetic ICU patients. These examples highlight AI's broad applicability in healthcare.

Emerging AI technologies are also advancing mental health diagnostics and therapy. Cho et al. [54] developed a ML-based mood prediction algorithm, offering new frameworks for clinical applications in mood disorders. AI models enhanced by UMLS and deep CNNs have proven effective in cancer detection [52], [53], mental health diagnosis [55], and diabetes complication management [56]. Additionally, Rasmy et al. [57] integrated logistic regression and recurrent neural networks (RNNs) with UMLS to predict heart failure risks in diabetic patients, showcasing AI's growing role in managing chronic diseases.

3 PROPOSED APPROACH

After conducting a thorough analysis and reviewing previous studies that address the various RQs in our systematic review, we have identified that modern technologies such as AI, big data, DL, and ML have the capacity to significantly enhance healthcare outcomes. Our findings suggest that the key to achieving accurate predictions of medical pathologies within an optimal timeframe lies in implementing a clear and reliable process. This process encompasses all stages leading up to the prediction phase, beginning with the collection of medical data, followed by pre-processing, mapping the processed data, and then classifying and clustering patient profiles. These steps are crucial for making precise pathology predictions within a defined time period. Based on these insights, we propose a comprehensive approach to improve both the accuracy and timeliness of clinical pathology predictions, organized around five critical steps:

3.1 Collecting data

The first step of our proposed approach (see Figure 2) involves aggregating medical data, such as electronic medical records and doctors' notes, in various formats *(PDF, RTF, HTML, XML)*. A *"Collection"* function is established, which processes the input file directory path *(DME)* and identifies both the format and location of each file. Additionally, a *"DATA"* class is created to store the path and format of each file.

To address the complexity of large-scale medical data, a *"Data Analyzer"* function is introduced, which takes the *"DATA"* class as input and categorizes the files along two primary axes:

- **Text (RTF, PDF):** Comprises all unstructured files that require more extensive analysis and processing.
- **Relational DB (.csv, .db):** Contains semi-structured files that demand less intensive analysis.

A series of wrappers is employed to execute sub-queries across the various data sources and convert the results into JSON documents, which are then stored in the *"Data Set"*.

At the end of this data collection phase, a *"Parser"* is applied, taking the data from the Data Set as input and producing the following output:

- **Meta data:** This includes critical information about each file, facilitating easier searching and archiving.
- **The new "EMR" dataset:** This dataset encompasses all the essential fields required for processing and analyzing patient medical data, which helps streamline data mining, reduce complexity, and save time by focusing only on relevant data for analysis.

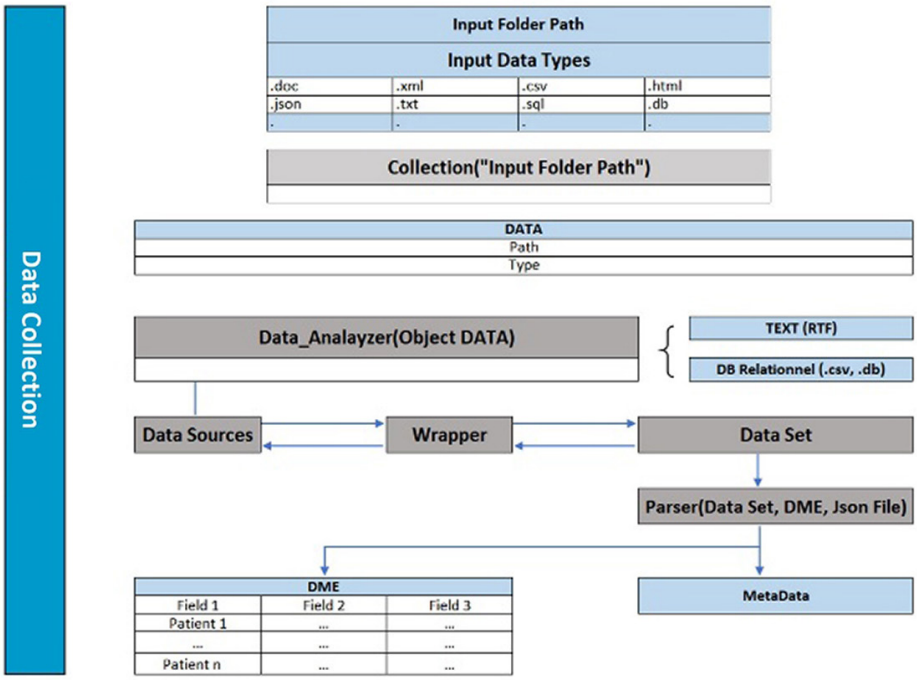

Fig. 2. Collecting data

3.2 Preprocessing data

The second step of our proposed approach (see Figure 3) focuses on applying the pre-processing phase to the *"DME"* dataset. This phase involves tasks such as data cleaning, integration, reduction, and transformation of the medical data using the NLTK [58], a Python-based library distributed under the general public license (GPL). NLTK provides a collection of modules, datasets, and tutorials designed to support the study and teaching of computational linguistics and NLP. Key features include transparent syntax, effective string handling, and simplicity. Utilizing NLTK's Treebank word tokenizer and POS tagger, the preprocessing phase achieves high performance, resulting in a structured and pre-processed dataset known as *"SDME."*

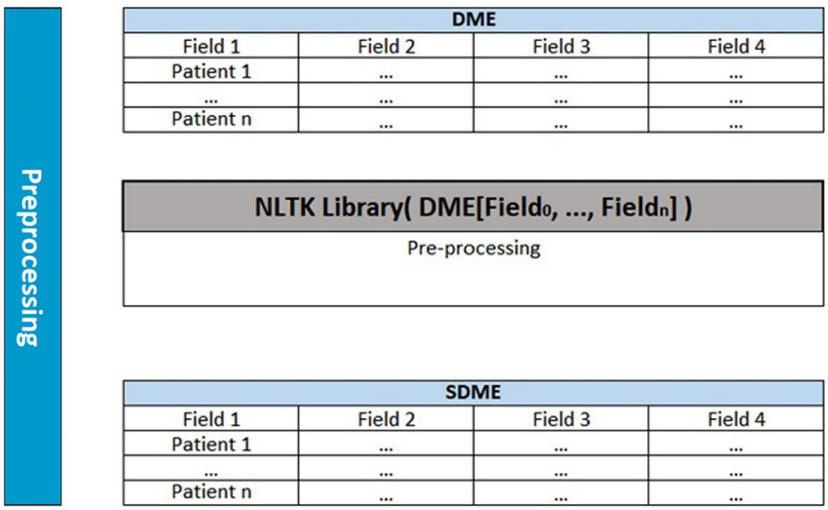

Fig. 3. Pre-processing data

3.3 Mapping medical data

The third step of our approach (see Figure 4) involves mapping biomedical terms within the structured *"SDME"* dataset using tools from the UMLS. The UMLS Meta thesaurus is a comprehensive repository of biomedical terminologies, integrating multiple health and medical vocabularies such as systematized nomenclature of medicine-clinical terms (SNOMED CT), International classification of diseases (ICD), and medical subject headings (MeSH). This facilitates mapping diverse terms to standardized concepts, ensuring semantic interoperability and harmonization across heterogeneous data sources.

To enhance the mapping process, we employ the GATE Bio-YODIE system, a robust biomedical entity recognition tool that performs NER and disambiguation. This system identifies various biomedical entities within the text and associates them with the most relevant conceptual tags from the UMLS Metathesaurus. This process ensures accurate extraction of biomedical terms and their alignment with established medical standards.

Furthermore, the mapped terms adhere to UMLS-based standards, which are applied to maintain consistency and interoperability across datasets. These standards leverage the semantic network within UMLS, including 135 semantic types and 54 relationships, to classify and categorize concepts. This ensures that the resulting data is not only standardized but also ready for downstream processing and analysis. The output of this step is the *"Mapped SDME"* dataset, which includes the following fields:

- **Mandatory terms:** A list of words or symptoms, each mapped to the UMLS Meta thesaurus for precise terminology alignment.
- **Concept extraction:** Concepts representing the meanings of medical terms, with each concept linked to its corresponding standardized identifiers.
- **Semantic type extraction:** This field leverages the UMLS semantic network to classify and categorize the extracted concepts.
- **Entity type extraction:** Displays the hierarchical relationships of the concepts, presenting their standardized meanings in an accessible format for downstream tasks.

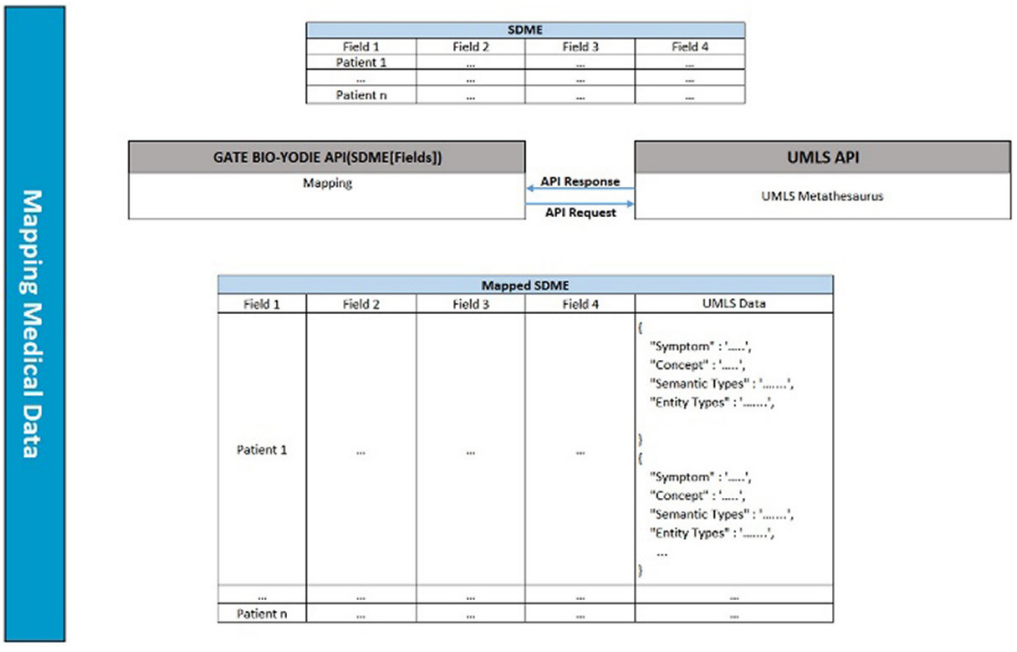

Fig. 4. Mapping medical data

3.4 Classification and clustering

The fourth step of our proposed approach (see Figure 5) involves the classification and clustering of patient profiles. A *"Clustering"* function is introduced, which processes the data from the *"Mapped SDME"* dataset as input and applies the K-means clustering algorithm. This widely used data mining technique divides large datasets into *k* distinct clusters, grouping objects based on their similarity to predefined class criteria [59].

In the proposed framework, K-means clustering plays a pivotal role in stratifying patients based on the mapped data, enabling the identification of homogeneous subgroups within heterogeneous datasets. By analyzing the data derived from the *UMLS Data* field, the K-means algorithm assigns each patient to one or more clusters. This facilitates the grouping of patients with similar clinical profiles, which is critical for tailoring personalized treatment plans and improving diagnostic precision.

The output of this step is a new dataset called *"Clustered Mapped SDME,"* which includes a *"Cluster"* field that defines the cluster(s) associated with each patient. This intermediate result serves as a foundation for subsequent predictive modeling by enhancing the interpretability and structure of the data.

Additionally, the simplicity, scalability, and computational efficiency of K-means make it ideal for handling large-scale healthcare data, ensuring robust performance even in scenarios with substantial data heterogeneity. By integrating this clustering step, the framework addresses key challenges in patient data management, enabling better insights into disease progression and facilitating targeted interventions.

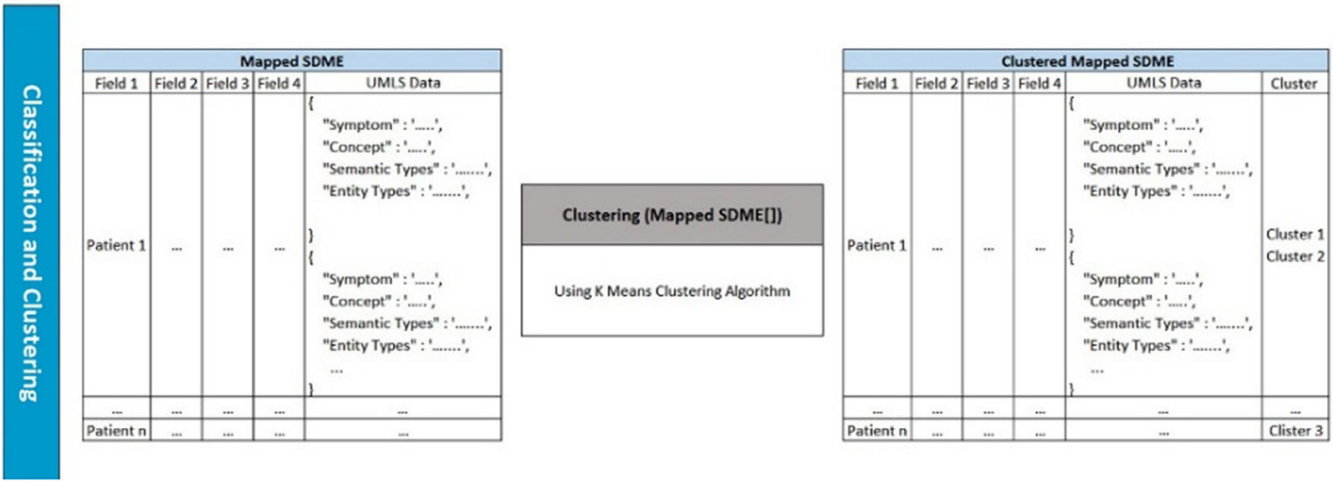

Fig. 5. Classification and clustering

3.5 Pathology prediction

The final step of our approach (see Figure 6) involves creating an AI model using the RNN algorithm. RNNs are particularly well-suited for handling sequential data because they utilize loops and memory components, enabling them to preserve information from earlier computations, which differentiates them from traditional feedforward neural networks [60].

In the proposed framework, RNNs are employed to process the data from the *"Clustered Mapped SDME"* dataset, capturing temporal dependencies and patterns

in medical histories. This capability allows the model to predict disease progression and identify critical events based on patient-specific sequences of data, such as longitudinal EHR entries and time-series diagnostic results. By leveraging this sequential modeling capability, RNNs enhance the framework's ability to provide precise and clinically relevant pathology predictions. The output of this step is the "*Pathology Prediction*" dataset, which includes three key fields:

- **"Pathologies" field:** Predicts the most frequent and severe pathologies.
- **"Best Prediction" field:** Allows for more precise pathology predictions.
- **"Best Precision" field:** Enables predictions to be made within a significant and relevant timeframe.

By integrating RNNs into the framework, the proposed approach effectively addresses critical challenges in healthcare. It supports accurate modeling of temporal data while generating actionable predictions in dynamic and time-sensitive clinical contexts. This enhancement strengthens diagnostic efficiency, improves predictive accuracy, and promotes personalized healthcare interventions.

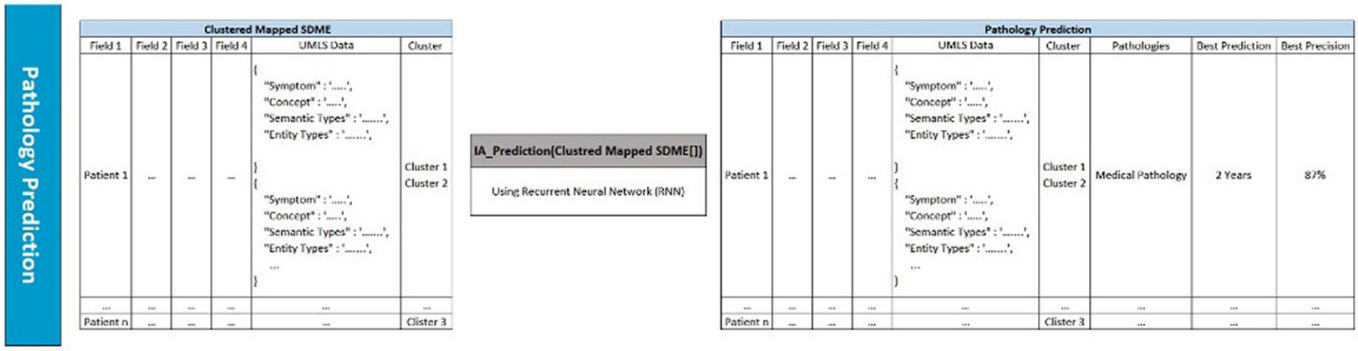

Fig. 6. Pathology prediction

4 RESULTS AND DISCUSSION

This systematic review identified several advanced techniques and methods employed in medical informatics for data collection, preprocessing, mapping, classification, and pathology prediction. These methods leverage innovative technologies, including big data, ML, and AI, to enhance healthcare outcomes. The results provide insights into the current state of research and highlight the advantages of the proposed approach in addressing existing challenges.

Process mining in healthcare is a growing field, highlighting the importance of domain experts in ensuring that insights are clinically relevant and applicable. Addressing the complexities of healthcare challenges requires interdisciplinary collaboration [61]. This aligns with the findings of the systematic review, which emphasize the necessity of integrating expertise across disciplines to develop effective solutions.

Table 2 presents a consolidated comparison of key findings from the systematic review and the proposed approach, emphasizing its ability to integrate and improve upon current methodologies. By combining structured and unstructured datasets, utilizing automated preprocessing, and leveraging advanced ML techniques, the proposed approach addresses critical gaps identified in the literature.

Table 2. Comparison of scientific approaches and proposed approach

Category	Current Scientific Approaches	Proposed Approach	Advantages of Proposed Approach
Data Collection	EHR and big data techniques [9], [10]	Combines structured and unstructured data for better preparation	Reduces complexity, enhances data processing efficiency
Preprocessing	NLP for entity recognition and relation extraction [11], [13]	Automated preprocessing with NLTK	Ensures better data structuring, improving downstream analysis
Data Mapping	UMLS for medical term mapping [24], [28]	UMLS and Bio-YODIE for enhanced concept extraction	Accurately identifies biomedical entities for better interpretability
Classification	Machine learning (Random Forest, Logistic Regression) [33], [44]	K-means clustering for patient stratification	Provides better insights for targeted patient interventions
Pathology Prediction	CNN and RNN for disease prediction [43], [52]	RNN-based pathology prediction	Ensures accurate and timely predictions

4.1 Key insights

The systematic review revealed that while existing approaches demonstrate significant potential, they often lack the flexibility to handle heterogeneous datasets and face limitations in scalability and interpretability. For instance, traditional ML models such as random forests lack adaptability for unstructured datasets, whereas neural networks require significant computational power. The proposed framework bridges these gaps by integrating preprocessing and mapping tools, enabling seamless handling of diverse datasets. The use of RNN for sequential modeling further enhances prediction accuracy, particularly in time-sensitive medical scenarios.

4.2 Limitations and challenges

While the proposed framework offers significant potential in addressing key challenges in healthcare informatics, certain limitations must be considered. Collecting large-scale healthcare data, particularly unstructured data from diverse institutions, remains a challenge due to issues related to privacy, data standardization, and availability. Variations in data quality and completeness can further affect model performance.

Additionally, many healthcare systems rely on legacy software and fragmented data storage architectures, making the integration of advanced AI frameworks difficult. Ensuring interoperability between different systems and data formats is another critical hurdle. Furthermore, implementing the framework demands substantial computational resources for pre-processing, mapping, and model training, which could limit its adoption in resource-constrained settings, such as rural or developing healthcare environments.

Overcoming these challenges will require further research into optimizing algorithms, developing lightweight AI models, and fostering collaboration between healthcare providers and technology developers to ensure seamless integration and practical applicability.

5 CONCLUSION AND FUTURE DIRECTIONS

This study provides a systematic review of recent advancements in AI, big data, DL, and ML techniques applied to medical pathology prediction. Drawing insights from 61 key studies, we proposed a novel framework that integrates structured and unstructured data with advanced pre-processing and predictive modeling techniques. By addressing critical gaps such as data fragmentation, limited generalizability, and interpretability issues, the framework demonstrates significant potential to enhance diagnostic efficiency and accuracy in healthcare.

However, the real-world implementation of the proposed approach presents certain challenges. Testing the framework on large-scale, real-world datasets will be essential to evaluate its performance using standardized metrics such as precision, recall, and F1-score. Furthermore, hybrid models that combine rule-based systems with ML techniques could significantly improve interpretability, addressing concerns about the "black-box" nature of AI algorithms. These hybrid approaches could balance the need for accuracy with the demand for explainability, ensuring that clinicians can trust and effectively utilize AI-based tools.

Emerging technologies, such as federated learning and edge computing, offer promising directions for decentralizing data processing. These approaches could mitigate computational constraints and enable the adoption of AI-driven solutions in resource-constrained environments, such as rural or developing healthcare systems. Additionally, tailoring the framework to address domain-specific challenges, such as rare diseases or regional healthcare systems, could further enhance its relevance and impact across diverse contexts.

By overcoming these challenges and embracing innovative perspectives, the proposed framework has the potential to pave the way for more precise, scalable, and accessible medical solutions, ultimately transforming the landscape of healthcare informatics.

6 REFERENCES

[1] K. Benke and G. Benke, "Artificial intelligence and big data in public health," *Int. J. Environ. Res. Public Health*, vol. 15, no. 12, p. 2796, 2018. https://doi.org/10.3390/ijerph15122796

[2] S. Dash *et al.*, "Big data in healthcare: Management, analysis and future prospects," *J. Big Data*, vol. 6, 2019. https://doi.org/10.1186/s40537-019-0217-0

[3] B. S. Glicksberg *et al.*, "PatientExploreR: An extensible application for dynamic visualization of patient clinical history from electronic health records in the OMOP common data model," *Bioinformatics*, vol. 35, no. 21, pp. 4515–4518, 2019. https://doi.org/10.1093/bioinformatics/btz409

[4] T. J. Callahan, I. J. Tripodi, H. Pielke-Lombardo, and L. E. Hunter, "Knowledge-based biomedical data science," *Annu. Rev. Biomed. Data Sci.*, vol. 3, pp. 23–41, 2020. https://doi.org/10.1146/annurev-biodatasci-010820-091627

[5] A. Tayal, J. Gupta, A. Solanki, K. Bisht, A. Nayyar, and M. Masud, "DL-CNN-based approach with image processing techniques for diagnosis of retinal diseases," *Multimedia Syst.*, vol. 28, pp. 1417–1438, 2022. https://doi.org/10.1007/s00530-021-00769-7

[6] J. Konttila *et al.*, "Healthcare professionals' competence in digitalisation: A systematic review," *J. Clin. Nurs.*, vol. 28, nos. 5–6, pp. 745–761, 2019. https://doi.org/10.1111/jocn.14710

[7] Y. Mintz and R. Brodie, "Introduction to artificial intelligence in medicine," *Minim. Invasive Ther. Allied Technol.*, vol. 28, no. 2, pp. 73–81, 2019. https://doi.org/10.1080/13645706.2019.1575882

[8] K. Seol, Y. G. Kim, E. Lee, Y. D. Seo, and D. K. Baik, "Privacy-preserving attribute-based access control model for XML-based electronic health record system," *IEEE Access*, vol. 6, pp. 9114–9128, 2018. https://doi.org/10.1109/ACCESS.2018.2800288

[9] R. H. Hariri, E. M. Fredericks, and K. M. Bowers, "Uncertainty in big data analytics: Survey, opportunities, and challenges," *J. Big Data*, vol. 6, pp. 1–16, 2019. https://doi.org/10.1186/s40537-019-0206-3

[10] J. Munoz-Gama *et al.*, "Process mining for healthcare: Characteristics and challenges," *J. Biomed. Inform.*, vol. 127, p. 103994, 2022.

[11] S. Modi *et al.*, "Extracting adverse drug events from clinical notes: A systematic review of approaches used," *J. Biomed. Inform.*, vol. 151, p. 104603, 2024. https://doi.org/10.1016/j.jbi.2024.104603

[12] D. Yogish, T. N. Manjunath, and R. S. Hegadi, "Review on natural language processing trends and techniques using NLTK," in *Recent Trends in Image Processing and Pattern Recognition, RTIP2R 2018, Communications in Computer and Information Science*, K. Santosh and R. Hegadi, Eds., Springer, Singapore, vol. 1037, 2018, pp. 589–606. https://doi.org/10.1007/978-981-13-9187-3_53

[13] N. K. Jha, "An approach towards text to emoticon conversion and vice-versa using NLTK and WordNet," in *Proc. 2018 2nd Int. Conf. Data Sci. Bus. Analytics (ICDSBA),* 2018, pp. 161–166. https://doi.org/10.1109/ICDSBA.2018.00036

[14] J. Yao, "Automated sentiment analysis of text data with NLTK," *J. Phys. Conf. Ser.*, vol. 1187, no. 5, pp. 1–8, 2019. https://doi.org/10.1088/1742-6596/1187/5/052020

[15] J. Wang *et al.*, "Systematic evaluation of research progress on natural language processing in medicine over the past 20 years: Bibliometric study on PubMed," *J. Med. Internet Res.*, vol. 22, no. 1, p. e1681, 2020. https://doi.org/10.2196/16816

[16] M. Assale *et al.*, "The revival of the notes field: Leveraging the unstructured content in electronic health records," *Front. Med.*, vol. 6, 2019. https://doi.org/10.3389/fmed.2019.00066

[17] V. Carchiolo, A. Longheu, G. Reitano, and L. Zagarella, "Medical prescription classification: A NLP-based approach," in *Proc. 2019 Fed. Conf. Comput. Sci. Inf. Syst. (FedCSIS),* vol. 18, 2019, pp. 605–609. https://doi.org/10.15439/2019F197

[18] H. Pooja and M. P. Prabhudev Jagadeesh, "A collective study of data mining techniques for the big health data available from the electronic health records," in *Proc. 2019 1st Int. Conf. Adv. Technol. Intell. Control, Environ. Comput. Commun. Eng. (ICATIECE),* 2019, pp. 51–55. https://doi.org/10.1109/ICATIECE45860.2019.9063623

[19] R. P. Reddy, C. Mandakini, and C. Radhika, "A review on data mining techniques and challenges in medical field," *Int. J. Eng. Res. Technol. (IJERT)*, vol. 9, no. 8, pp. 329–333, 2020. https://doi.org/10.17577/IJERTV9IS080143

[20] G. Ganino, D. Lembo, M. Mecella, and F. Scafoglieri, "Ontology population for open-source intelligence: A GATE-based solution," *Softw.: Pract. Exp.*, vol. 48, no. 12, pp. 2302–2330, 2018. https://doi.org/10.1002/spe.2640

[21] A. Amjad and U. Qamar, "UAMSA: Unified approach for multilingual sentiment analysis using GATE," in *Proc. 6th Conf. Eng. Comput. Based Syst. (ECBS'19),* 2019, pp. 1–5. https://doi.org/10.1145/3352700.3352725

[22] V. C. Pezoulas *et al.*, "Medical data quality assessment: On the development of an automated framework for medical data curation," *Comput. Biol. Med.*, vol. 107, pp. 270–283, 2019. https://doi.org/10.1016/j.compbiomed.2019.03.001

[23] S. B. Goldberg *et al.*, "Machine learning and natural language processing in psychotherapy research: Alliance as example use case," *J. Counsel. Psychol.*, vol. 67, no. 4, pp. 438–448, 2020. https://doi.org/10.1037/cou0000382

[24] M. Topaz *et al.*, "Mining fall-related information in clinical notes: Comparison of rule-based and novel word embedding-based machine learning approaches," *J. Biomed. Inform.*, vol. 90, p. 103103, 2019. https://doi.org/10.1016/j.jbi.2019.103103

[25] A. Al-Hroob, A. T. Imam, and R. Al-Heisa, "The use of artificial neural networks for extracting actions and actors from requirements document," *Inf. Softw. Technol.*, vol. 101, pp. 1–15, 2018. https://doi.org/10.1016/j.infsof.2018.04.010

[26] Y. Wang *et al.*, "Clinical information extraction applications: A literature review," *J. Biomed. Inform.*, vol. 77, pp. 34–49, 2018. https://doi.org/10.1016/j.jbi.2017.11.011

[27] M. C. Kim, S. Nam, F. Wang, and Y. Zhu, "Mapping scientific landscapes in UMLS research: A scientometric review," *J. Am. Med. Inform. Assoc.*, vol. 27, no. 10, pp. 1612–1624, 2020. https://doi.org/10.1093/jamia/ocaa107

[28] G. Gorrell *et al.*, "Bio-yodie: A named entity linking system for biomedical text," *arXiv preprint arXiv:1811.04860*, 2018.

[29] A. Abbas *et al.*, "Explicit and implicit section identification from clinical discharge summaries," in *2022 16th Int. Conf. Ubiquitous Inf. Manage. Commun. (IMCOM),* 2022, pp. 1–8. https://doi.org/10.1109/IMCOM53663.2022.9721771

[30] C. Taoussi, I. Hafidi, A. Metrane, and A. Lasbahani, "Predicting psychological pathologies from electronic medical records," in *Human Interaction, Emerging Technologies and Future Applications IV, IHIET-AI 2021,* in Advances in Intelligent Systems and Computing, T. Ahram, R. Taiar, and F. Groff, Eds., Springer, Cham, vol. 1378, 2021, pp. 493–500. https://doi.org/10.1007/978-3-030-74009-2_63

[31] M. Alawad, S. S. Hasan, J. B. Christian, and G. Tourassi, "Retrofitting word embeddings with the UMLS metathesaurus for clinical information extraction," in *2018 IEEE Int. Conf. Big Data (Big Data),* 2018, pp. 2838–2846. https://doi.org/10.1109/BigData.2018.8621999

[32] M. P. Maurits *et al.*, "A framework for employing longitudinally collected multicenter electronic health records to stratify heterogeneous patient populations on disease history," *J. Am. Med. Inform. Assoc.*, vol. 29, no. 5, pp. 761–769, 2022. https://doi.org/10.1093/jamia/ocac008

[33] C. Ricciardi *et al.*, "Classifying different stages of Parkinson's disease through random forests," in *XV Mediterranean Conference on Medical and Biological Engineering and Computing – MEDICON 2019, IFMBE Proceedings*, J. Henriques, N. Neves, and P. de Carvalho, Eds., Springer, Cham, vol. 76, 2019, pp. 1155–1162. https://doi.org/10.1007/978-3-030-31635-8_140

[34] Y. Wang *et al.*, "Unsupervised machine learning for the discovery of latent disease clusters and patient subgroups using electronic health records," *J. Biomed. Inform.*, vol. 102, p. 103364, 2020. https://doi.org/10.1016/j.jbi.2019.103364

[35] A. I. Kadhim, "An evaluation of preprocessing techniques for text classification," *Int. J. Comput. Sci. Inf. Secur. (IJCSIS)*, vol. 16, no. 6, pp. 22–32, 2018.

[36] M. Kashina, I. D. Lenivtceva, and G. D. Kopanitsa, "Preprocessing of unstructured medical data: The impact of each preprocessing stage on classification," *Procedia Comput. Sci.*, vol. 178, pp. 284–290, 2020. https://doi.org/10.1016/j.procs.2020.11.030

[37] L. Jerlin Rubini and E. Perumal, "Efficient classification of chronic kidney disease by using multi-kernel support vector machine and fruit fly optimization algorithm," *Int. J. Imaging Syst. Technol.*, vol. 30, no. 3, pp. 660–673, 2020. https://doi.org/10.1002/ima.22406

[38] L. Huang, A. L. Shea, H. Qian, A. Masurkar, H. Deng, and D. Liu, "Patient clustering improves efficiency of federated machine learning to predict mortality and hospital stay time using distributed electronic medical records," *J. Biomed. Inform.*, vol. 99, p. 103291, 2019. https://doi.org/10.1016/j.jbi.2019.103291

[39] H. Desai, "Biomedical data classification with improvised deep learning architectures," 2020.

[40] A. Wood, V. Shpilrain, K. Najarian, and D. Kahrobaei, "Private naive bayes classification of personal biomedical data: Application in cancer data analysis," *Comput. Biol. Med.*, vol. 105, pp. 144–150, 2019. https://doi.org/10.1016/j.compbiomed.2018.11.018

[41] M. Naegelin *et al.*, "An interpretable machine learning approach to multimodal stress detection in a simulated office environment," *J. Biomed. Inform.*, vol. 139, p. 104299, 2023. https://doi.org/10.1016/j.jbi.2023.104299

[42] A. A. Aouragh, M. Bahaj, and F. Toufik, "Diabetes prediction: Optimization of machine learning through feature selection and dimensionality reduction," *Int. J. Online Eng. (iJOE)*, vol. 20, no. 8, pp. 100–114, 2024. https://doi.org/10.3991/ijoe.v20i08.47765

[43] S. J. MacEachern and N. D. Forkert, "Machine learning for precision medicine," *Genome*, vol. 64, no. 4, pp. 416–425, 2021. https://doi.org/10.1139/gen-2020-0131

[44] M. Cabanillas-Carbonell and J. Zapata-Paulini, "Improving the accuracy of oncology diagnosis: A machine learning-based approach to cancer prediction," *Int. J. Online Eng. (iJOE)*, vol. 20, no. 11, pp. 102–122, 2024. https://doi.org/10.3991/ijoe.v20i11.49139

[45] H. Vega-Huerta, K. R. Pantoja-Pimentel, S. Y. Quintanilla-Jaimes, G. L. E. Maquen-Niño, P. De-La-Cruz-VdV, and L. Guerra-Grados, "Classification of Alzheimer's disease based on deep learning using medical images," *Int. J. Online Eng. (iJOE)*, vol. 20, no. 10, pp. 101–114, 2024. https://doi.org/10.3991/ijoe.v20i10.49089

[46] L. Muflikhah, A. G. Nurfansepta, F. A. Bachtiar, and D. E. Ratnawati, "High performance for predicting diabetic nephropathy using stacking regression of ensemble learning method," *Int. J. Online Eng. (iJOE)*, vol. 20, no. 8, pp. 149–164, 2024. https://doi.org/10.3991/ijoe.v20i08.48387

[47] D. Dahiwade, G. Patle, and E. Meshram, "Designing disease prediction model using machine learning approach," in *2019 3rd Int. Conf. Comput. Methodol. Commun. (ICCMC)*, 2019, pp. 1211–1215. https://doi.org/10.1109/ICCMC.2019.8819782

[48] M. E. Shipe, S. A. Deppen, F. Farjah, and E. L. Grogan, "Developing prediction models for clinical use using logistic regression: An overview," *J. Thorac. Dis.*, vol. 11, no. Suppl. 4, pp. S574–S584, 2019. https://doi.org/10.21037/jtd.2019.01.25

[49] S. Nusinovici *et al.*, "Logistic regression was as good as machine learning for predicting major chronic diseases," *J. Clin. Epidemiol.*, vol. 122, pp. 56–69, 2020. https://doi.org/10.1016/j.jclinepi.2020.03.002

[50] F. Ramzan *et al.*, "A deep learning approach for automated diagnosis and multi-class classification of Alzheimer's disease stages using resting-state fMRI and residual neural networks," *J. Med. Syst.*, vol. 44, 2020. https://doi.org/10.1007/s10916-019-1475-2

[51] J. Ye, L. Yao, J. Shen, R. Janarthanam, and Y. Luo, "Predicting mortality in critically ill patients with diabetes using machine learning and clinical notes," *BMC Med. Inform. Decis. Mak.*, vol. 20, no. Suppl. 11, pp. 1–7, 2020. https://doi.org/10.1186/s12911-020-01318-4

[52] M. Alawad *et al.*, "Retrofitting word embeddings with the UMLS metathesaurus for clinical information extraction," in *2018 IEEE Int. Conf. Big Data (Big Data)*, 2018, pp. 2838–2846. https://doi.org/10.1109/BigData.2018.8621999

[53] S. Wang *et al.*, "Comprehensive analysis of lung cancer pathology images to discover tumor shape and boundary features that predict survival outcome," *Sci. Rep.*, vol. 8, pp. 1–9, 2018. https://doi.org/10.1038/s41598-018-27707-4

[54] C. H. Cho, T. Lee, M. G. Kim, H. P. In, L. Kim, and H. J. Lee, "Mood prediction of patients with mood disorders by machine learning using passive digital phenotypes based on the circadian rhythm: Prospective observational cohort study," *J. Med. Internet Res.*, vol. 21, no. 4, p. e11029, 2019. https://doi.org/10.2196/11029

[55] S. G. Alonso *et al.*, "Data mining algorithms and techniques in mental health: A systematic review," *J. Med. Syst.*, vol. 42, 2018. https://doi.org/10.1007/s10916-018-1018-2

[56] B. Ljubic *et al.*, "Predicting complications of diabetes mellitus using advanced machine learning algorithms," *J. Am. Med. Inform. Assoc.*, vol. 27, no. 9, pp. 1343–1351, 2020. https://doi.org/10.1093/jamia/ocaa120

[57] L. Rasmy *et al.*, "Representation of EHR data for predictive modeling: A comparison between UMLS and other terminologies," *J. Am. Med. Inform. Assoc.*, vol. 27, no. 10, pp. 1593–1599, 2020. https://doi.org/10.1093/jamia/ocaa180

[58] S. Bird, "NLTK: The natural language toolkit," in *Proc. COLING/ACL 2006 Interactive Presentation Sessions*, 2006, pp. 69–72. https://doi.org/10.3115/1225403.1225421

[59] H. K. Bharadwaj *et al.*, "A review on the role of machine learning in enabling IoT based healthcare applications," *IEEE Access*, vol. 9, pp. 38859–38890, 2021. https://doi.org/10.1109/ACCESS.2021.3059858

[60] A. Pramod, H. S. Naicker, and A. K. Tyagi, "Machine learning and deep learning: Open issues and future research directions for the next 10 years," in *Computational Analysis and Deep Learning for Medical Care: Principles, Methods, and Applications*, 2021, pp. 463–490. https://doi.org/10.1002/9781119785750.ch18

[61] E. De Roock and N. Martin, "Process mining in healthcare–An updated perspective on the state of the art," *J. Biomed. Inform.*, vol. 127, p. 103995, 2022. https://doi.org/10.1016/j.jbi.2022.103995

7 AUTHORS

Chaimae Taoussi is a PhD student in computer science and applied mathematics at the Laboratory of Process Engineering, Computer Science, and Mathematics (LIPIM) at Sultan Moulay Slimane University. Currently, in the fourth year of the PhD program, the research focuses on the prediction of medical pathologies using big data and artificial intelligence. The work aims to utilize advanced data analytics and machine learning techniques to enhance medical diagnostics and improve healthcare outcomes. Research interests include AI, big data, computer science, and healthcare informatics (E-mail: chaimae.taoussi@usms.ac.ma).

Imad Hafidi is currently a Professor at the National School of Applied Science (ENSA) in Khouribga, part of Sultan Moulay Slimane University. He is the Head of the Department of Mathematics and Computer Engineering and the Director of the Laboratory of Process Engineering, Computer Science, and Mathematics (LIPIM) at ENSA Khouribga. His research interests include wireless sensor networks (WSN), machine learning, big data, and computer vision (E-mail: i.hafidi@usms.ma).

Abdelmoutalib Metrane is currently a Professor at the Faculty of Sciences and Technology (FST) in Marrakech, part of Cadi Ayyad University. Previously, he served as a professor and Head of the Department of Mathematics and Computer Engineering at the National School of Applied Sciences (ENSA) in Khouribga. He pursued his studies at Polytechnique Montréal. His research interests include machine learning, big data, operations research, logistics, and production management (E-mail: a.metrane@uca.ac.ma).