Natural Language Processing with Deep Learning

AE1 **-** Word documentation

Student ID: 23220052

Part1: General Questions (15 Marks):

1. Provide examples of Natural Language Processing (NLP) systems and describe the primary functions typically associated with an NLP application.

Natural Language Processing (NLP) are systems built on interaction between computers and humans through natural language, enabling them to understand, interpret, analyse and generate human language efficiently [1]. These systems perform various core functions such as text parsing and tokenization which involve breaking down text into smaller components like words or sentences to facilitate analysis such as email filtering [2]. Another core function is sentiment analysis which consists of determining the text sentiment, as seen in IBM Watson Natural Language Understanding, which analyses social media posts or reviews to determine the sentiment expressed. Moreover, Named Entity Recognition (NER) consists of identifying and categorising key information (name, dates and location) within the text, such as with spaCy. Furthermore, NLPs also include machine translation such as Google Translate, enabling real-time language translation to facilitate communication. Also, text summarisation such as OpenAI's GPT model generates short summaries for quick information retrieval. Finally, speech recognition and synthesis such as Google Speech-to-Text, Siri and Alexa convert spoken language into text, enabling voice typing, transcription services, voice commands and virtual assistants [3].

References:

1. IBM (2021). What Is NLP (Natural Language Processing)? | IBM. [online] Available at:https://www.ibm.com/topics/natural-language-processing#:~:text=NLP%20enables%20computers%20and%20digital.
2. Tableau (2022). 8 Natural Language Processing (NLP) Examples. [online] Tableau. Available at: https://www.tableau.com/learn/articles/natural-language-processing-examples.
3. MonkeyLearn. (n.d.). Natural Language Processing (NLP): What Is It & How Does it Work? [online] Available at: https://monkeylearn.com/natural-language-processing/#:~:text=Natural%20Language%20Processing%20(NLP)%20allows.

2. Explain the role of RNNs or LSTMs in NLP for handling sequential data. Provide examples of NLP tasks where these architectures are effective and discuss their advantages compared to scenarios where RNNs or LSTMs might not be the best choice and alternative architectures that could be more suitable.

Being a class of Neural Networks (NN), Recurrent Neural Networks (RNNs) are deep learning models commonly used in speech recognition and NLPs [1]. They are trained to process and convert sequential data one element at a time, maintaining an internal state that is influenced by previous inputs. This makes them effective for sequential data tasks as they recognize data's sequential characteristics, and use patterns to predict the next likely scenario. By maintaining an internal memory, they can capture temporal dependencies and patterns through loops in their architecture [2]. These RNNs are usually found in applications such as Siri, voice search, and Google Translate as they are used for tasks such as language translation, speech recognition, and image captioning. Similarly, Long Short-Term Memory (LSTM) networks are an improved version of RNNs, that better retain and manage data over long sequences [2]. Indeed, they excel in sequence prediction tasks, capturing long-term dependencies, which makes them well-suited for tasks like language translation, speech recognition, and time series forecasting [3].

These systems are highly effective in language translation as they handle long sequences and maintain context over extended sentences, as seen in tools like Google Translate which leverage LSTMs for accurate and fluent translations [4]. Moreover, they excel in speech recognition as they process audio signals sequentially, and understand temporal patterns which is crucial for accurately converting spoken language into text, as used in applications like Siri and voice search [5]. Furthermore, in text generation tasks, such as predictive text input or chatbot responses, they generate coherent and contextually appropriate text by learning from large datasets, predicting the next word based on previous context, and maintaining context over long sequences to create human-like text [5]. Finally, in terms of sentiment analysis, RNNs are effective in understanding the context of words in a sentence and the sentiment expressed, resulting in more accurate sentiment classification in reviews or social media posts [5].

However, despite their powerful capabilities, RNNs face limitations with very long sequences, high computational costs, and capturing global context, making training challenging and inefficient [6,7]. Therefore, alternative systems like Transformers, CNNs, and BERT are more suitable for tasks requiring extensive context understanding, as they capture dependencies between all sequence elements regardless of distance, offering better performance and scalability for large-scale NLP tasks, such as language modelling, machine translation, text summarization, and question answering [8]. Indeed, transformers handle long-range dependencies and process sequences simultaneously using self-attention, CNNs are faster and more effective for tasks like text classification, and BERT captures context from both directions, enhancing performance in tasks requiring a comprehensive understanding of global context [7].

References:

1. IBM (2023). What are Recurrent Neural Networks? | IBM. [online] www.ibm.com. Available at: https://www.ibm.com/topics/recurrent-neural-networks.
2. Donges, N. (2019). Recurrent neural networks 101: Understanding the basics of RNNs and LSTM. [online] Built In. Available at: https://builtin.com/data-science/recurrent-neural-networks-and-lstm.
3. Amazon Web Services, Inc. (n.d.). What is RNN? - Recurrent Neural Networks Explained - AWS. [online] Available at: https://aws.amazon.com/what-is/recurrent-neural-network/.
4. Badrinarayan, M. (2024). Sequence-to-Sequence Models for Language Translation. [online] Analytics Vidhya. Available at: https://www.analyticsvidhya.com/blog/2024/05/sequence-to-sequence-models-for-language-translation/ [Accessed 18 Jul. 2024].
5. Chugh, A. (2019). Deep Learning | Introduction to Long Short Term Memory. [online] GeeksforGeeks. Available at: https://www.geeksforgeeks.org/deep-learning-introduction-to-long-short-term-memory/.
6. Indrajitbarat (2023). Recurrent Neural Networks (RNNs): Challenges and Limitations. [online] Medium. Available at: https://medium.com/@indrajitbarat9/recurrent-neural-networks-rnns-challenges-and-limitations-4534b25a394c [Accessed 18 Jul. 2024].
7. Nicole Laskowski, (2021.). What are recurrent neural networks and how do they work? [online] Available at: https://www.techtarget.com/searchenterpriseai/definition/recurrent-neural-networks.
8. Choi, S.R. and Lee, M. (2023). Transformer Architecture and Attention Mechanisms in Genome Data Analysis: A Comprehensive Review. Biology, [online] 12(7), p.1033. doi:https://doi.org/10.3390/biology12071033.

3. Explain the concept of distributed representations in word embeddings. How does it differ from traditional one-hot encoding?

Distributed representations in word embeddings is a method that groups similar words and captures their semantic meaning based on their context [1]. Indeed, each word is mapped to high-dimensional vectors in a continuous vector space, allowing similar words to have similar representations [2]. In other words, this method converts categorical data into continuous numerical vectors, where words that occur in similar contexts are represented by a pattern of values across many dimensions and are positioned close to each other. These distributed representations have a wide range of applications, such as measuring word similarity, text classification, machine translation, information retrieval, and sentiment analysis [3]. Learning these representations involves training models on tasks that capture semantic or feature similarities, such as predicting a word from its context or predicting surrounding words from a given word. One popular method for training word embeddings is

Word2Vec, which uses an NN to predict the surrounding words of a target word in a given context [4, 5]. Another approach is GloVe, which leverages global statistics to create embeddings [5]. However, distributed representations require large amounts of data to learn meaningful relationships, are computationally intensive, and lack interpretability [6]. In contrast, one-hot encoding is a traditional method where words are represented as sparse binary vectors with little to no information about their relationships to other words and a dimension equal to the size of the vocabulary [6, 7]. Despite being simple and intuitive, this method does not capture any semantic relationships between words [7].

These two methods differ in terms of representation, dimensionality, semantic meaning, static vs. learned representations, and NLP task performance. Indeed, one-hot encoding represents each word as a sparse vector with one element set to 1, whereas distributed representations use dense vectors of real numbers, capturing semantic and syntactic relationships between words based on their contexts. Regarding the dimensionality, one-hot encoding requires high-dimensional vectors equal to the vocabulary size, while word embeddings use low-dimensional vectors (typically 50 to 300 dimensions), making them more memory-efficient and computationally manageable [8]. In terms of semantic meaning, one-hot encoding represents each word independently without capturing relationships between words, whereas word embeddings ensure that similar words have similar vector representations. Furthermore, one-hot encoding provides static representations based on word positions in the vocabulary, while word embeddings are learned through training on large text datasets to better capture semantic and syntactic relationships [8]. Finally, in terms of use in NLP tasks, one-hot encoding is limited due to its inability to capture semantic relationships and its high-dimensional nature, whereas distributed representations are widely used in NLP tasks such as text classification, machine translation, and sentiment analysis, by providing richer semantic representations of words [8]. However, choosing the best approach depends on the specific problem and dataset. Indeed, it is more useful to use one-hot encoding with small datasets while embeddings are often the better choice for larger ones [9].

References:

1. DeepAI. (2019). Distributed Representations. [online] Available at: https://deepai.org/machine-learning-glossary-and-terms/distributed-representation [Accessed 18 Jul. 2024].
2. Uma, C. (2024) *What is text embedding for ai? transforming NLP with ai*, *DataCamp*. Available at: https://www.datacamp.com/blog/what-is-text-embedding-ai (Accessed: 18 July 2024).
3. IBM (n.d.). What are Word Embeddings? | IBM. [online] Available at: https://www.ibm.com/topics/word-embeddings.
4. Sharvil (2020). Word Embeddings Deep Dive — Hands-on approach. [online] Medium. Available at: https://towardsdatascience.com/word-embeddings-deep-dive-hands-on-approach-a710eb03e4c5 [Accessed 18 Jul. 2024].
5. Durna, M.B. (2024). Advanced Word Embeddings: Word2Vec, GloVe, and FastText. [online] Medium. Available at: https://medium.com/@mervebdurna/advanced-word-embeddings-word2vec-glove-and-fasttext-26e546ffedbd.

6. Elastic (n.d.). What are Word Embeddings? | A Comprehensive Word Embedding Guide. [online] Available at: https://www.elastic.co/what-is/word-embedding.
7. Prasan, N. H. (2024). Text Representation: One-Hot Encoding - Prasan N H - Medium. [online] Medium. Available at: https://medium.com/@prasanNH/text-representation-one-hot-encoding-e18257434395#:~:text=One%20commonly%20used%20method%20for [Accessed 18 Jul. 2024].
8. Kumar , S. (2024) *Understanding differences between encoding and embedding*, *LinkedIn*. Available at: https://www.linkedin.com/pulse/understanding-differences-between-encoding-embedding-mba-ms-phd/ (Accessed: 18 July 2024).
9. Tanvir (2020). Word Embedding and One hot encoding. [online] intelligentmachines. Available at: https://medium.com/intelligentmachines/word-embedding-and-one-hot-encoding-ad17b4bbe111.

Part 2: ***Coding Questions (55 Marks):***

Text Normalisation, Language Modeling, and Text Classification:

1. Explain the concept of distributed representations in word embeddings. How does it differ from traditional one-hot encoding?

In order to ensure that our text is preprocessed effectively, it is vital to prioritize normalization steps. Indeed, in our case, we first remove any URL from our text as they do not contribute to any semantic meaning and therefore ensure that they are not being processed in our tokenization and lemmatization step. Next, we convert our text to lowercase to ensure uniformity and matching words accurately as well as prevent case sensitivity issues (such as duplication of words based on case differences) during our tokenization and lemmatization steps. After that, we tokenize our text into sentences to maintain their context. It is important to note that this step relies on punctuation, so it should precede its removal to avoid disrupting the process and causing inaccurate segmentation of text into sentences. Then, we remove any punctuation and symbols as they do not carry semantic meaning on their own, to ensure that we have clean tokens and to prevent symbols from being treated as separate tokens. Please, note that removing symbols first ensures that numerical values remain intact and are correctly recognized during subsequent processing steps. After that, we do the same with numbers as they do not carry any semantic meaning independently. This is done to simplify our text, reduce the complexity of the vocabulary, avoid unnecessary tokens and improve model performance by focusing on textual contents. However, in our text, as numbers are quite significant, we realised that it would have been better if we converted them to categorical data for better representation. Following this, we tokenize our sentences into words, making our text prepared for lemmatization which consists of reducing the words into their base forms to ensure that they are in their standard, normalized form, which is crucial for consistency and analysis.

2. e) Compare the results and performance of the two designed classifiers.

Our results from Naive Bayes and BiLSTM classifiers indicate a strong performance across key evaluation metrics. Indeed, for our Naive Bayes model, the accuracy, precision, recall, and F1-score are all approximately 83%. This suggests that our model performance is consistent in predicting both positive and negative classes in our dataset. On the other hand, the BiLSTM model shows a slightly higher performance metrics with an accuracy of 86.65%, and a precision, recall and F1-score of 86%. These results indicate that our BiLSTM model outperforms Naive Bayes in terms of overall predictive accuracy and the ability to correctly identify and classify instances of both classes. Indeed, being a more complex NN architecture capable of capturing sequential dependencies in data, our BiLSTM model is better suited for tasks where context and the order of words matter, such as in NLP. On the other side, while being simpler and easier to interpret, Naive Bayes shows a slightly lower performance metrics but still maintains strong overall predictive capability. Therefore, the choice between these two models will depend on specific application requirements, balancing interpretability against higher predictive accuracy and data complexity.

***Part 3:** Task: Technical Report on NLP Applications (30 Marks):*

Topic: 2. Explainable AI in NLP: Interpretable Models and Ethical Considerations
Discussion Question: What are the implications of using explainable AI techniques in NLP for ensuring transparency and accountability?
Word count: 2254

I - Introduction:

Natural Language Processing (NLP) is a field of artificial intelligence that focuses on the interaction between computers and human through natural language, enabling them to understand, interpret, analyse and generate human language efficiently [1]. However, with the growing need of algorithms transparency and interpretability, Explainable AI (XAI) has become critical due to the models increasing complexity and opacity [2]. Indeed, XAI refers to the methods that make the outputs of machine learning (ML), deep learning (DL) and NLP models understandable to humans [3]. In NLPs, these powerful models, often function as "black boxes," making it difficult to understand how they reach specific conclusions. XAI aims to clarify the AI system's hidden mechanisms, providing clear explanations for their decisions. This new method consists of various techniques, such as attention mechanisms, post-hoc interpretation methods model, interpretability and visualization, designed to improve NLP system reliability, AI models and their decision-making processes [3] by highlighting areas of potential bias or limitations. By providing insights into AI system's

behaviour and ensuring their results are transparent and interpretable, XAI plays a crucial role in fields requiring high transparency, trust and accountability such as healthcare. Indeed, AI-driven systems are increasingly used for tasks like diagnosing diseases, predicting patient outcomes, and personalising treatment plans [4]. However, the opaque nature of many AI models can be a significant barrier to their widespread adoption in clinical settings [5]. By incorporating XAI, healthcare professionals (HCPs) can gain insights into AI's decision-making processes allowing them to trust AI's recommendations, leading to better patient care and adherence to ethical standards [4].

⇒ In this report, we will first shed light on explainable AI techniques and challenges in NLP. Then, we will demonstrate its impact on the healthcare industry, and finally discuss the implications of using explainable AI techniques in NLP to ensure transparency and accountability.

II - Techniques:

XAI in NLP uses several methods to improve the transparency and understandability of decision-making processes. These techniques collectively promote transparency and accountability in NLP applications ensuring that users can trust and comprehend AI systems [2].

Feature importance and attribution methods, such as LIME (Local Interpretable Model-agnostic Explanations) and SHAP (SHapley Additive exPlanations), are commonly employed to identify which features (words, phrases) in a text most influence a model's predictions or output [6]. These methods assess the significance of features offering insights into the models decision making process even though they may be computationally intensive and struggle with complex models like transformers [7]. For example, in cardiovascular disease diagnosis, these models might identify terms such as "chest pain," "shortness of breath," "elevated blood pressure," and "history of smoking" as critical indicators. By highlighting these terms, HCPs gain clear insights into what drives the models decisions thereby boosting transparency and confidence in AI driven diagnoses [8].

Moreover, attention mechanisms, particularly in transformer models, inherently provide some level of interpretability by indicating which parts of the input text the model focuses on during processing [7]. Visualisation of attention weights, often via heatmaps, helps users understand the model's interpretation of the text, although these weights do not always correspond perfectly to human-interpretable importance [9]. On one side, model distillation involves training a simpler, more interpretable model to mimic the behaviour of a complex model [10]. This strategy simplifies complex models into more interpretable ones, such as linear models or decision trees, providing clearer explanations of the decision-making process. While this enhances understandability, it may reduce accuracy and fail to capture the full complexity of the original model [11]. On the other side, counterfactual explanations involve modifying input data to observe how it affects the output, providing actionable insights into the decision-making process [12]. This can be done by altering specific words or

phrases in a text to observe how these changes impact the model prediction. However, generating realistic and meaningful counterfactuals can be challenging [13]. Finally, post-hoc analysis techniques, like counterfactual explanations, involve slight alterations of inputs to understand the decision boundaries of the model and the causal relationships between inputs and predictions [14].

For instance, Payrovnaziri et al. (2020) applied XAI techniques to predict patient readmission risks using electronic health records (EHRs) [15]. Their model combined LSTM networks with attention mechanisms to generate a risk score for readmission by examining patients demographic data, medical history and previous healthcare patterns [16] to predict the likelihood of a patient being readmitted within 30 days of discharge. The attention mechanism highlighted important features in the EHRs, such as specific diagnoses, treatments, and patient demographics that significantly contributed to the prediction. Additionally, SHAP values quantified the impact of each feature, providing a clear explanation of why a particular prediction was made. Whereas counterfactual analysis was also conducted to understand how slight changes in patient data could affect readmission predictions, offering insights into the model's decision boundaries [15].

III- Challenges:

However, XAI in NLPs faces several significant challenges.
Modern NLP models are highly complex and difficult to interpret. Indeed, the techniques designed to provide interpretability often struggle to scale effectively with the size and complexity of these models, leading to trade-offs between interpretability and performance[17]. Despite achieving high accuracy, complex algorithms with million of parameters based on ML, DL and NLPs become increasingly integrated into decision-making processes [18], making their mechanism difficult for humans to understand [19]. However, their main issue resides in the lack of transparency regarding the rationale behind their decision-making, which raises concerns about accountability, bias and trustworthiness [20]. For example, in the healthcare industry, these systems can be used to facilitate disease diagnosis and treatment recommendation [21], but they produce an outcome with no process justification causing clinicians to be reluctant to use their outcome [22]. Since these decisions impact patients directly, HCPs need to understand how these decisions were made to avoid accountability and trust concerns in technology adoption. For instance, in medical imaging, CNNs process vast amounts of data and extract complex interactions between features and weights with high accuracy [23], which may not be evident to clinicians. However, their complexity prevents clinicians from interpreting the features driving these predictions, making them hesitant to follow the resulting recommendation blindly [24]. Given that the fundamental focus of these algorithms is on accuracy and performance rather than interpretability, their lack of transparency leads to HCPs' reluctance to adopt AI-driven solutions in clinical practice, limiting their potential impact and benefits.

Secondly, what constitutes an interpretable explanation can be subjective and context-dependent, varying with the stakeholders' requirements and explanation preferences [3]. An explanation that satisfies one stakeholder might not meet the needs of

another, highlighting the challenge of developing XAI techniques that are universally accepted and effective across all stakeholder groups [3]. For example, when predicting the likelihood of patients developing diabetes based on their EHRs, data scientists might focus on the model's technical accuracy and the statistical significance of features like blood glucose levels and BMI [25]. Whereas HCPs might require explanations that are easy to understand and actionable, such as identifying lifestyle factors or patient behaviours that can be modified to reduce risk. On the other hand, regulatory bodies such as the European Union's General Data Protection Regulation (GDPR), might prioritise transparency and the model's adherence to privacy and ethical standards, ensuring that no sensitive patient information is improperly used or exposed [26]. Indeed, they have established standards that require transparency within medical AI for clinical applications [27] to uphold ethical standards and promote society's best interests. Therefore, this stakeholder requirement variability underscores the challenge of creating explanations that accommodate the diverse contexts and requirements of different stakeholders.

Thirdly, interpretability alone does not ensure fairness or ethical decision-making. Explanations can inadvertently expose or amplify biases present in the data or model, necessitating careful consideration of the data, context, and potential impacts on different user groups to ensure ethical AI deployment [3]. In healthcare datasets, which usually contain diverse patient demographics, clinical and medical imaging data [28], handling such an extensive and diverse dataset increases algorithm complexity as well as the lack of transparency. Their complexity can hinder comprehension, accuracy in decision-making, clinical validation, and correction methods, limiting treatment recommendation understanding and potentially leading to therapeutic misdirection [29]. Similarly, if these systems are trained on biased data, they can amplify biases in their recommendations, resulting in wrong predictions and undermining error detection credibility, thus affecting disease prediction, diagnosis, and treatment recommendations [30]. Therefore, without transparency, clinicians may be unaware of algorithm limitations, struggle to understand their decision-making process and fail to identify and mitigate biases, compromising patient safety and potentially leading to invalid or discriminatory outcomes [31]. For instance, analysing NGS data is a laborious process that consists of comparing the generated data with a known reference genome to detect mutations [32]. However, training datasets may be biased due to their limited diversity. This results in reference genomes that may display biases towards specific populations, leading to discrepancies in sequence alignment accuracy for underrepresented groups. As per Dr. Asmann statement, “The reference genome itself has racial bias because it came from patients mostly of European ancestry" [33]. These biases can distort gene function predictions, impacting the understanding of gene expression data. Consequently, predictive models trained on biased data, generate wrong predictions, affecting the credibility of error detection and leading to potentially inaccurate outcomes in disease prediction, diagnosis, and treatment recommendations.

IV - Discussion:

The adoption of XAI techniques in NLP holds significant implications for ensuring transparency and accountability across various domains.

Firstly, XAI enhances user trust by providing understandable explanations for model decisions, thereby increasing acceptance of outcomes in critical scenarios such as medical diagnoses [3]. This transparency is crucial in healthcare settings, where HCPs rely on accurate predictions for patient care [34]. By documenting algorithms, ensuring code transparency, and tracking data provenance, XAI enhances transparency and instils HCPs' confidence in AI-driven recommendations over time [35]. Indeed, techniques like feature importance quantification and visualization methods such as heatmaps allow HCPs to grasp the key factors influencing AI predictions, facilitating validation and refinement of treatment recommendations [36]. Moreover, by demystifying complex algorithms, XAI empowers HCPs to interpret and validate decisions effectively, thus promoting wider adoption of AI technologies in clinical practice [4]. For example, the complexity of AI algorithms, particularly in fields like Next-Generation Sequencing (NGS) for gene expression analysis, poses challenges for decision-making, clinical validation, and correction methods, potentially leading to concerns about transparency [37]. However, according to Mittelstadt et al., (2019), prioritising transparency in algorithmic decision-making and ensuring clear explanations for visual representations and predictive models are essential steps forward [38]. Indeed, interactive interpretability models, proposed as a solution by Mittelstadt et al., (2019) engage users in exploring input changes and their impact on output predictions, enhancing understanding of clinical diagnoses and risk factors [38]. Moreover, integrating genetic testing into eHRs through Natural Language Generation (NLG) facilitates the conversion of raw genetic data into user-friendly reports, improving communication and enabling precision medicine [39]. Furthermore, implementing rigorous validation procedures and transparent reporting of algorithm performance enhances comprehension of genomic information, thereby reinforcing trust in clinical decision-making and treatment options.

Secondly, by offering insights into decision-making processes, XAI techniques play a crucial role in ensuring regulatory compliance, as it helps healthcare organisations meet regulatory standards to enhance transparency and accountability. Indeed, regulatory agencies such as the European Union's GDPR mandates explanations for automated decisions [27], the adherence to patient privacy, data security, and informed consent [40], as well as the collaboration among HCPs, computer scientists, ethicists, and policymakers to develop ethical XAI frameworks that protect patient privacy, ensure fairness, and enhance transparency in decision-making processes to protect individual and societal interests [41].

Finally, XAI is crucial to detect and mitigate biases in AI systems by identifying influential features in predictions, which is critical for applications in healthcare where biased models can disproportionately impact patient outcomes [42]. Therefore, standardized evaluation metrics for interpretability and transparency are essential for benchmarking AI systems and ensuring their effectiveness across diverse patient populations and healthcare settings [43] can enhance their transparency, trustworthiness, and effectiveness [44]. Moreover, involving HCPs in algorithm design and testing phases, where they provide feedback, identify and mitigate biases, review model predictions, and make adjustments to optimise patient outcomes and quality of care, improves the accuracy and reliability of diagnosis and treatment recommendations [35, 45].

V - Conclusion:

In conclusion, the integration of XAI in NLP holds immense potential for enhancing transparency, trust, and accountability, particularly in the healthcare sector. As AI systems grow in complexity and opacity, XAI becomes essential for elucidating their decision-making processes, making their outputs understandable and trustworthy. Indeed, the adoption of XAI in NLP represents a transformative step toward more transparent, reliable, and ethical AI systems. By addressing the challenges of interpretability, bias, and regulatory compliance, XAI ensures that AI technologies can be confidently and effectively integrated into critical domains like healthcare, ultimately benefiting patients and society at large. However, XAI techniques also present challenges in terms of scalability, usability, and effectiveness. Nevertheless, the implications of XAI extend beyond technical transparency to include enhanced user trust, regulatory compliance, bias detection, and improved model development. Yet, requiring complete transparency and interpretability, is often impractical, particularly in highly complex AI systems. The vast data volume and complexity processed by these systems make it challenging, if not impossible, to clearly explain every decision or prediction they generate. By excessively simplifying these systems, we risk compromising their ability to address complex problems. Moreover, generating full interpretability may hinder innovation in AI research and development, as strict interpretability requirements might discourage researchers from exploring novel, more complex techniques that could lead to significant advancements in AI capabilities.

VI - References:

1) IBM (2021). What Is NLP (Natural Language Processing)? | IBM. [online] Available at:https://www.ibm.com/topics/natural-language-processing#:~:text=NLP%20enables%20computers%20and%20digital.
2) Saeed, W. and Omlin, C. (2023). Explainable AI (XAI): A systematic meta-survey of current challenges and future opportunities. Knowledge-Based Systems, 263, p.110273. doi:https://doi.org/10.1016/j.knosys.2023.110273.
3) Ali, S., Abuhmed, T., El-Sappagh, S., Muhammad, K., Alonso-Moral, J.M., Confalonieri, R., Guidotti, R., Ser, J.D., Díaz-Rodríguez, N. and Herrera, F. (2023). Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence. Information Fusion, [online] 99(101805), p.101805. doi:https://doi.org/10.1016/j.inffus.2023.101805.
4) Alowais, S.A., Alghamdi, S.S., Alsuhebany, N., Alqahtani, T., Alshaya, A., Almohareb, S.N., Aldairem, A., Alrashed, M., Saleh, K.B., Badreldin, H.A., Yami, A., Harbi, S.A. and Albekairy, A.M. (2023). Revolutionizing healthcare: the role of artificial intelligence in clinical practice. BMC Medical Education, [online] 23(1). doi:https://doi.org/10.1186/s12909-023-04698-z.
5) Molla Imaduddin Ahmed, Spooner, B., Isherwood, J., Lane, M.A., Orrock, E. and Dennison, A. (2023). A Systematic Review of the Barriers to the Implementation of Artificial Intelligence in Healthcare. Cureus, 15(10). doi:https://doi.org/10.7759/cureus.46454.
6) Aas, K., Jullum, M. and Løland, A. (2021). Explaining individual predictions when features are dependent: More accurate approximations to Shapley values. Artificial Intelligence, p.103502. doi:https://doi.org/10.1016/j.artint.2021.103502.

7) Choi, S.R. and Lee, M. (2023). Transformer Architecture and Attention Mechanisms in Genome Data Analysis: A Comprehensive Review. Biology, [online] 12(7), p.1033. doi:https://doi.org/10.3390/biology12071033.
8) Wysocki, O., Davies, J.K., Vigo, M., Armstrong, A.C., Landers, D., Lee, R. and Freitas, A. (2022). Assessing the communication gap between AI models and healthcare professionals: Explainability, utility and trust in AI-driven clinical decision-making. Artificial Intelligence, 316, p.103839. doi:https://doi.org/10.1016/j.artint.2022.103839.
9) Mohamed, E., Sirlantzis, K. and Howells, G. (2022). A review of visualisation-as-explanation techniques for convolutional neural networks and their evaluation. Displays, 73, p.102239. doi:https://doi.org/10.1016/j.displa.2022.102239.
10) labelbox.ghost.io. (n.d.). A pragmatic introduction to model distillation for AI developers. [online] Available at: https://labelbox.com/blog/a-pragmatic-introduction-to-model-distillation-for-ai-developers/ [Accessed 18 Jul. 2024].
11) Science direct (n.d.). Knowledge Distillation - an overview | ScienceDirect Topics. [online] Available at: https://www.sciencedirect.com/topics/computer-science/knowledge-distillation [Accessed 18 Jul. 2024].
12) KPMG. (n.d.). Counterfactual Explanations: The What-Ifs of AI Decision Making. [online] Available at: https://kpmg.com/ch/en/insights/technology/artificial-intelligence-counterfactual-explanation.html#:~:text=Counterfactual%20explanations%20are%20essential%20in [Accessed 18 Jul. 2024].
13) Javier Del Ser, Barredo-Arrieta, A., Díaz-Rodríguez, N., Herrera, F., Saranti, A. and Holzinger, A. (2024). On generating trustworthy counterfactual explanations. Information sciences, 655, pp.119898–119898. doi:https://doi.org/10.1016/j.ins.2023.119898.
14) Retzlaff, C.O., Angerschmid, A., Saranti, A., Schneeberger, D., Röttger, R., Müller, H. and Holzinger, A. (2024). Post-hoc vs ante-hoc explanations: xAI design guidelines for data scientists. Cognitive Systems Research, [online] 86, p.101243. doi:https://doi.org/10.1016/j.cogsys.2024.101243.
15) Payrovnaziri, S.N. *et al.* (2020) *Explainable artificial intelligence models using real-world electronic health record data: A systematic scoping review*, *Journal of the American Medical Informatics Association : JAMIA*. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7647281/ (Accessed: 19 July 2024).
16) Mahmoudi, E., Kamdar, N., Kim, N., Gonzales, G., Singh, K., & Waljee, A. K. (2020). Use of electronic medical records in development and validation of risk prediction models of hospital readmission: systematic review. The BMJ, 369. https://doi.org/10.1136/bmj.m958
17) Jaotombo, F et al., (2023). Finding the best trade-off between performance and interpretability in predicting hospital length of stay using structured and unstructured data. PLOS ONE, [online] 18(11), pp.e0289795–e0289795. doi:https://doi.org/10.1371/journal.pone.0289795.
18) Iqbal, J., Cortés Jaimes, D. C., Makineni, P., Subramani, S., Hemaida, S., Thugu, T. R., Butt, A.N., Sikto, J. T., Kaur, P., Lak, M. A., Augustine, M., Shahzad, R., & Arain, M.

(2023). Reimagining Healthcare: Unleashing the Power of Artificial Intelligence in Medicine. Cureus. https://doi.org/10.7759/cureus.44658
19) Censius.ai. (n.d.). What is AI Black Box - Responsible AI | MLOps Wiki. [online] Available at: https://censius.ai/wiki/ai-black-box [Accessed 19 July. 2024].
20) Osasona, F et al., (2024). REVIEWING THE ETHICAL IMPLICATIONS OF AI IN DECISION MAKING PROCESSES. International Journal of Management & Entrepreneurship Research. 6. 322-335. 10.51594/ijmer.v6i2.773
21) Chan, B. (2023). Black-box assisted medical decisions: AI power vs. ethical physician care. Medicine, Health Care and Philosophy, 26(3). https://doi.org/10.1007/s11019-023-10153-z
22) Ahmed, M. I., Spooner, B., Isherwood, J., Lane, M., Orrock, E., & Dennison, A. (2023). A Systematic Review of the Barriers to the Implementation of Artificial Intelligence in Healthcare. Cureus. https://doi.org/10.7759/cureus.46454
23) Agrawal , K. (2023). Revolutionizing Brain Tumor Detection: The CNN-Based Medical Imaging Breakthrough. [online] INFORMS.org. Available at: https://doi.org/10.1287/orms.2023.04.03.
24) Sarvamangala, D. R., & Kulkarni, R. v. (2022). Convolutional neural networks in medical image understanding: a survey. In Evolutionary Intelligence (Vol. 15, Issue 1). https://doi.org/10.1007/s12065-020-00540-3
25) Mohsen, F., Al-Absi, H.R.H., Yousri, N.A., El Hajj, N. and Shah, Z. (2023). A scoping review of artificial intelligence-based methods for diabetes risk prediction. npj Digital Medicine, [online] 6(1), pp.1–15. doi:https://doi.org/10.1038/s41746-023-00933-5.
26) Pool, J.K. et al., (2024). A systematic analysis of failures in protecting personal health data: A scoping review. International Journal of Information Management, [online] 74(102719), pp.102719–102719. doi:https://doi.org/10.1016/j.ijinfomgt.2023.102719.
27) Reddy, S. (2023). Navigating the AI Revolution: The Case for Precise Regulation in Health Care. Journal of Medical Internet Research, 25(1). https://doi.org/10.2196/49989
28) Ehrenstein, V., Kharrazi, H., Lehmann, H. and Taylor, C.O. (2019). Obtaining Data From Electronic Health Records. [online] www.ncbi.nlm.nih.gov. Agency for Healthcare Research and Quality (US). Available at: https://www.ncbi.nlm.nih.gov/books/NBK551878/.
29) Yang, G., Ye, Q., & Xia, J. (2022). Unbox the black-box for the medical explainable AI via multi-modal and multi-centre data fusion: A mini-review, two showcases and beyond. Information Fusion, 77. https://doi.org/10.1016/j.inffus.2021.07.016
30) Hassija, V., Chamola, V., Mahapatra, A., Singal, A., Goel, D., Huang, K., Scardapane, S., Spinelli, I., Mahmud, M., & Hussain, A. (2024). Interpreting Black-Box Models: A Review on Explainable Artificial Intelligence. In Cognitive Computation (Vol. 16, Issue 1). https://doi.org/10.1007/s12559-023-10179-8
31) Jain, A., Brooks, J. R., Alford, C. C., Chang, C. S., Mueller, N. M., Umscheid, C. A., & Bierman, A.S. (2023). Awareness of Racial and Ethnic Bias and Potential Solutions to Address Bias With Use of Health Care Algorithms. JAMA Health Forum, 4(6). https://doi.org/10.1001/jamahealthforum.2023.1197
32) Dahui, Q. (2019) 'Next-generation sequencing and its clinical application', National Library of Medicine, 16(1), pp. 4–10. doi:10.20892/j.issn.2095-3941.2018.0055

33) Mayo Clinic (2022) Racial disparities discovered among genomic sequencing data . Available at: https://www.mayoclinic.org/medical-professionals/news/racial-disparities-discovered-among-genomic-sequencing-data/mac-20534960 (Accessed: 19 July F 2024).
34) Al Kuwaiti, A., Nazer, K., Al-Reedy, A., Al-Shehri, S., Al-Muhanna, A., Subbarayalu, A.V., Al Muhanna, D. and Al-Muhanna, F.A. (2023). A Review of the Role of Artificial Intelligence in Healthcare. Journal of Personalized Medicine, 13(6), p.951. doi:https://doi.org/10.3390/jpm13060951.
35) Tsamados, A., Aggarwal, N., Cowls, J., Morley, J., Roberts, H., Taddeo, M., & Floridi, L. (2022). The ethics of algorithms: key problems and solutions. AI and Society, 37(1). https://doi.org/10.1007/s00146-021-01154-8
36) Champendal M., Müller H., Prior, J.O. and dos, S. (2023). A Scoping Review of Interpretability and Explainability concerning Artificial Intelligence Methods in Medical Imaging. European Journal of Radiology, pp.111159–111159. doi:https://doi.org/10.1016/j.ejrad.2023.111159.
37) Satam, H. et al. (2023) 'Next-generation sequencing technology: Current trends and advancements', National Library of Medicine, 12(7), p. 997. doi:10.3390/biology12070997.
38) Mittelstadt, B., Russell, C. and Wachter, S. (2019) 'Explaining Explanations in AI', DL ACM [Preprint] doi:10.1145/3287560.3287574.
39) Wigmore, I. (2023) What is natural language generation (NLG)?: Definition from TechTarget, Enterprise AI. Available at: https://www.techtarget.com/searchenterpriseai/definition/natural-language-generation-NLG (Accessed: 26 February 2024).
40) Siala, H., & Wang, Y. (2022). SHIFTing artificial intelligence to be responsible in healthcare: A systematic review. Social Science and Medicine, 296 https://doi.org/10.1016/j.socscimed.2022.114782
41) Dimitra Panteli, Helena Legido-Quigley, Christoph Reichebner, Günter Ollenschläger, Corinna Schäfer, & Reinhard Busse. (2019). Clinical practice guidelines as a quality strategy. Improving Healthcare Quality in Europe: Characteristics, Effectiveness and Implementation of Different Strategies
42) Ueda, D., Kakinuma, T., Fujita, S., Kamagata, K., Fushimi, Y., Ito, R., Matsui, Y., Nozaki, T., Nakaura, T., Fujima, N., Tatsugami, F., Yanagawa, M., Hirata, K., Yamada, A., Tsuboyama, T., Kawamura, M., Fujioka, T. and Naganawa, S. (2023). Fairness of Artificial Intelligence in healthcare: Review and Recommendations. Japanese Journal of Radiology, 42(1). doi:https://doi.org/10.1007/s11604-023-01474-3.
43) Ray, P. P. (2023). Benchmarking, ethical alignment, and evaluation framework for conversational AI: Advancing responsible development of ChatGPT. BenchCouncil Transactions on Benchmarks, Standards and Evaluations, 3(3). https://doi.org/10.1016/j.tbench.2023.100136
44) Fehr, J., Citro, B., Malpani, R., Lippert, C., & Madai, V. I. (2024). A trustworthy AI reality-check: the lack of transparency of artificial intelligence products in healthcare. Frontiers in Digital Health, 6. https://doi.org/10.3389/fdgth.2024.1267290
45) Grote, T., & Berens, P. (2020). On the ethics of algorithmic decision-making in healthcare. In Journal of Medical Ethics (Vol. 46, Issue 3). https://doi.org/10.1136/medethics-2019-105586