Data Preprocessing For Machine Learning Applications in Healthcare: A Review

1st Ismail Chahid
LARI Laboratory
Faculty of Science, Mohammed First University
Oujda, Morocco
ismail.chahid@gmail.com

2nd Aissa Kerkour Elmiad
LARI Laboratory
Faculty of Science, Mohammed First University
Oujda, Morocco
mid.kerkour@gmail.com

3rd Mohammed Badaoui
LARI Laboratory
Faculty of Science, Mohammed First University
Oujda, Morocco
med.badaoui@gmail.com

2023 14th International Conference on Intelligent Systems: Theories and Applications (SITA) | 979-8-3503-0821-1/23/$31.00 ©2023 IEEE | DOI: 10.1109/SITA60746.2023.10373591

***Abstract*— Data preprocessing plays a critical role in the success of machine learning and deep learning models in the medical and healthcare field. As the availability of healthcare data continues to grow, ensuring its quality, reliability, and suitability for machine learning tasks becomes essential. In this paper, we will try to provide an in-depth exploration of data preprocessing techniques specifically tailored to the medical and healthcare domain. We will cover various steps involved in data preprocessing, including data types, data cleaning, data transforming, and data normalization. Additionally, challenges and considerations unique to medical data preprocessing are discussed.**

Keywords—Data Preprocessing, Machine Learning, Dataset.

I. INTRODUCTION

Recent artificial intelligence (AI) techniques are fundamentally transforming healthcare, ushering in a new era of possibilities. Natural language processing, deep learning, machine learning, and other cognitive technologies are being applied to improve disease prevention, treatment adherence [1], monitoring, rehabilitation, diagnosis, and predictive methods. These progressive AI approaches are fueling a global shift toward intelligent healthcare, bridging gaps in both affluent and underserved communities. The transition from in-person to remote care underscores AI's profound impact, enhancing healthcare delivery universally [2]. Intelligent chatbots, digital health tools, and applications, crafted through natural language processing, deep learning, and machine learning, are reshaping medical practices and services. These technologies have proven their worth in specialized fields like radiology, psychiatry, pathology, and ophthalmology, enabling early detection, prediction, diagnosis, and management of diverse health conditions, ranging from cancer and diabetes to tuberculosis and HIV/AIDS. AI-powered healthcare applications empower professionals and policymakers, equipping them to make insightful decisions, personalize patient care, and elevate overall healthcare services [3].

Machine learning (ML) and Deep learning (DL) falls under the domain of artificial intelligence (AI) and involves the utilization of models that acquire knowledge through training and/or experience [4]. ML can be categorized into four main branches: supervised learning, unsupervised learning, semi-supervised learning, and reinforcement learning. Supervised learning involves algorithms being provided with input examples along with their corresponding desired outputs, referred to as training data. The objective is to learn general rules that establish a mapping between inputs and outputs. The input data, known as training data, is accompanied by known outputs. This training data serves as a guide for the algorithm's development. Through a training process, a model is constructed, wherein predictions are made and corrected when they prove to be inaccurate. The training process continues until the model attains the desired level of accuracy on the training data. Unsupervised learning involves a learning algorithm that operates without receiving labeled data, thereby relying on the input to discover inherent structure. Semi-supervised learning is a combination approach that incorporates both labeled and unlabeled data in the input. While there may exist a desired outcome, the model also learns underlying structures to organize the data and make predictions. This approach is applicable to various problem types, such as classification and regression. Reinforcement learning represents a convergence of machine learning, behavioral psychology, ethics, and information theory. In this approach, the algorithm receives feedback in the form of rewards and punishments while navigating a given problem. By learning behaviors that maximize rewards, reinforcement learning enables the agent to determine the most favorable course of action based on its current state. The process of trial and error, combined with feedback, allows the algorithm to acquire an optimal policy or set of actions. This mechanism enables the algorithm to discern ideal behaviors within a given context.

The exponential growth of biomedical big data, stemming from health data collection through digital health wearables, genomic sequencing, and electronic health records (EHRs), presents another area of concern [5]. ML/DL approaches possess the capability to extract valuable insights from vast health datasets. Additionally, ML models contribute to enhancing care quality, patient safety, and overall cost efficiency in healthcare [6]. The extraction of pertinent data proves exceptionally advantageous in significantly addressing critical medical conditions. ML/DL approaches aid in extracting specific attributes and utilizing trained models for accurate diagnoses, prognoses, and interpretation of medical data and images [7]. This can facilitate the identification of high-risk patients, early detection of lung cancer, identification of abusive and fraudulent health insurance claims, and diagnosis of respiratory ailments through chest X-rays. However, regardless of the vast amount of data available and one's expertise in data science, the ability to derive meaningful insights from medical data records is essential for

the effectiveness of machine learning. Without this capability, a machine would be rendered nearly useless or even pose potential harm.

The reality is that all datasets possess inherent flaws. This is why data preprocessing holds immense significance in the machine learning process [8]. In essence, data preprocessing encompasses a series of procedures aimed at enhancing the suitability of the dataset for machine learning. This process also includes establishing the appropriate mechanisms for data collection. It is worth noting that these procedures often consume a substantial amount of time in the machine learning workflow, with months potentially passing before the first algorithm is developed.

II. DATA-PREPROCESSING

Data preprocessing is a critical step in the application of machine learning algorithms, particularly in the medical and healthcare field. The successful implementation of machine learning techniques relies heavily on the quality, relevance, and suitability of the data used for training and analysis **fig 1**.

In healthcare, where data is inherently complex, heterogeneous, and often incomplete, data preprocessing becomes even more crucial. This article explores the importance of data preprocessing in the medical and healthcare domain, focusing on its objectives, scope, and the various techniques and considerations that can enhance the outcomes of machine learning applications.

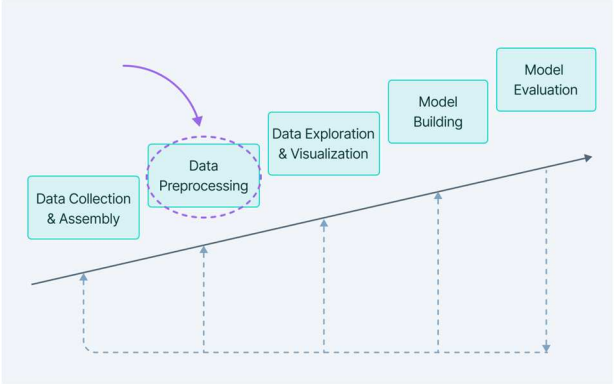

Fig. 1 : Data Preprocessing

Data preprocessing is of paramount importance in the healthcare industry due to the unique characteristics of medical data. Healthcare data is inherently complex, comprising diverse types such as electronic health records, medical imaging, genomics, wearable devices, and clinical trials. These data sources often exhibit significant heterogeneity, including variations in format, structure, and semantics. Moreover, medical data is frequently incomplete or contains missing values, making it necessary to address these gaps before utilizing the data for analysis. By performing data preprocessing, researchers and practitioners can ensure that the machine learning models are trained on high-quality, reliable data, leading to more accurate and meaningful results.

The objectives of data preprocessing in the medical and healthcare field are multifaceted, it aims to improve the quality of the data by identifying and addressing data quality issues, such as inconsistencies, errors, and outliers. Data cleaning techniques, such as removing duplicates, correcting errors, and handling missing values, are employed to ensure the integrity and accuracy of the data [9]. Data preprocessing involves feature selection, where relevant and informative features are identified and selected to enhance the performance and interpretability of machine learning models. Feature selection techniques help mitigate the curse of dimensionality and improve model efficiency. Data normalization techniques are employed to standardize and rescale the data, ensuring that different variables are on a comparable scale and eliminating bias caused by varying measurement units or ranges.

Various techniques and considerations are used for Data preprocessing in the medical and healthcare field. These include handling missing data [10], dealing with outliers, addressing noise, selecting relevant features, and extracting meaningful features. When it comes to missing data, imputation methods such as mean imputation, regression imputation, and multiple imputation are employed to estimate missing values and preserve the integrity of the data. Outliers, which can significantly impact the performance of machine learning models, are identified and treated using techniques such as z-score, interquartile range, or robust methods. Noise reduction techniques, such as filtering or smoothing, are applied to healthcare data to remove irrelevant or erroneous data points.

Feature selection techniques, including filter methods, wrapper methods, and embedded methods, are employed to identify the most relevant features for the predictive models. These techniques help reduce the dimensionality of the data, improve model interpretability, and enhance computational efficiency. In addition, feature extraction methods such as principal component analysis (PCA), independent component analysis (ICA), and non-negative matrix factorization (NMF) are utilized to capture the most informative features from high-dimensional healthcare data. These techniques facilitate dimensionality reduction while preserving relevant information.

1- Healthcare Data.

1-1 Structured Data

According to HIMSS (Healthcare Information and Management Systems Society), Structured data is data “organized into specific fields as part of a schema, with each field having a defined purpose". In healthcare databases, structured data encompasses various details such as customer names, contact information, lab values, patient demographic data, and financial information. This type of data is typically quantitative and can be readily formatted for databases, allowing for seamless integration into analytics and decision support systems.

1-2 Unstructured data

Unstructured data, as defined by HIMSS, refers to data that “cannot be easily organized using pre-defined structures.". In healthcare, examples of unstructured data include medical images, text files such as physician's notes in electronic health records (EHRs), as well as data from anonymous web users' comments, Twitter data, voice search data, and other online consumer touchpoints. Unlike structured data that can fit into predefined categories, unstructured data defies easy categorization, making it challenging to utilize without significant manual effort. While unstructured data can hold value, it requires structure to unlock its full potential.

1-3 Text data

Text data is individual words or sentences that can be utilized in various natural language processing (NLP) platforms, which enable computers to read, comprehend, and generate human language [11]. To make them compatible with machines, text data needs to be transformed into a numerical representation. In diagnostic surgical pathology, NLP integration can be applied to tasks like transcription or search. NLP faces challenges in addressing ambiguities such as homophones, syntax, semantics, and context. Homophones, for example, are challenging for machines to understand, particularly when considered without context. Understanding context is another important challenge in NLP. Our minds have developed a deep understanding of language usage in different scenarios and across various spoken languages, which machines lack. To compensate, machines rely on corpora [12], which are large collections of text data used to train and evaluate NLP algorithms. Translating complex language subtleties into code is difficult, emphasizing the importance of context and complicating the use of NLP models. However, newer NLP-based platforms and approaches, such as OpenAI's ChatGPT, are beginning to address these issues and making progress towards Artificial General Intelligence.

Text data preprocessing tasks can be divided into two categories: low-level and high-level tasks, and achieving the optimal NLP model relies on the synergistic interplay of these essential groups. Low-level tasks revolve around manipulating individual words or smaller language units and often involve processing raw text data. Tokenization is a crucial low-level task where text is split into individual words or units, enabling the machine to treat each instance as a distinct value, similar to numbers in tabular or image classification tasks. Removing stop words is another significant low-level task, as these commonly used words contribute little meaningful information on their own. Stemming, which reduces words to their base form, and part-of-speech tagging are additional low-level NLP tasks.

NLP plays a crucial role in preparing texts for processing by employing various steps, as outlined below:

- Phonological analysis: This step involves carefully selecting characters based on the sentence, following phonetic rules, phonemic rules, and prosodic rules. These rules govern the production sounds, pronunciations, and stress and intonation patterns in a sentence.
- Morphological analysis: This step examines the form of words in a language and their relationships, including inflectional and derivational analysis, which deal with terms having the same or different forms as their primary term, respectively.
- Lexical analysis: Also known as tokenization, this process extracts meaningful terms (tokens) from sentences, such as nouns, verbs, and adjectives.
- Syntactic analysis: Using a parser, this step verifies the accuracy of sentences and determines their language construction, including the role of each term, such as subjects, verbs, objects, and adjectives.
- Semantic analysis: This phase focuses on assessing the logical meaning of words using semantic networks that model word-level interactions. For example, it ensures that an apple cannot eat a mouse, but the opposite holds true.
- Pragmatic analysis: Given that terms can have multiple meanings depending on the context, this step considers the context of each sentence, paragraph, or entire document to accurately interpret the meaning of terms.

NLP techniques are typically employed during the pre-processing phase of text processing. The standard pre-processing operations for text involve the following steps:

- Tokenizing: This step separates meaningful terms, aligning with the lexical analysis in NLP steps.
- Stemming or lemmatization: This morphological analysis aims to find the base form of words, particularly for verbs or plural nouns with different forms.
- Removing stop words: These are irrelevant words, such as punctuation, conjunctions, copula verbs, and frequent terms whose presence and frequency in a context are not crucial. Eliminating stop words saves time and prevents misinterpretation of results.
- Typo detection and correction: Human-generated reports often contain typos, which can be quickly identified using NLP tools. Correcting typos, especially interconnected terms, improves text mining and information retrieval validation performance.

1-4 Image data

The data type referred to here consists of visual information, typically arranged as arrays of pixel values. In the field of pathology, an example of this would be microscopic images obtained from various tissue samples, such as bacteria shapes. These images can be used to train machine learning (ML) models that help differentiate between different types of images.

In order for these images to be compatible with machine learning libraries, they usually need to be organized into multidimensional arrays or tensors [13]. Each array represents specific characteristics of the image, such as its height and width, along with other intrinsic features. Additionally, there may be metadata attached to these files, providing additional information like the video frame number or specific location data.

To develop an optimized machine learning model, image preprocessing techniques like resizing and cropping are crucial. Resizing the images to a consistent dimension ensures that the neural network receives uniform inputs. Cropping allows for focused analysis, concentrating on a particular region of interest within the image. These techniques help eliminate noise and irrelevant information, thereby improving the accuracy of the model. Furthermore, other data augmentation techniques like flipping, rotating, and normalizing can be employed to increase the diversity of the training data. This leads to a more robust model that can generalize better to unseen data.

Image analysis can be performed either manually by a physician or automatically using computer programs. Manual analysis is often time-consuming, labor-intensive, and prone to errors, making automatic analysis preferable. Computer-

aided detection and diagnosis systems are equipped with modules capable of analyzing medical images. These modules employ intelligent algorithms like classification, regression, detection, and segmentation to carry out analytical tasks [14].

Classification algorithms categorize tumors, pathologies, or tissues into different sub-types.

Regression algorithms aim to assign numerical scores to images representing various risks or degrees of a specific disease. Detection algorithms focus on localizing pathologies and tissues of interest.

Segmentation algorithms accurately outline abnormalities. Image-based computer-aided systems are beneficial in screening programs as early disease diagnosis improves prognosis and reduces mortality rates.

1-5 Numerical and Tabular data

Numerical data in healthcare encompass measurements and counts of clinical, laboratory, or historical values, making them the most abundant type of data in the healthcare domain. Proper representation of numerical data ensures systematic and reproducible findings. These data can be classified as qualitative or categorical, contrasting with quantitative values **Fig2**.

PatientID	Gender	Age	Zip code	Test
55998	M	19	15723	Negative
88557	F	35	15674	Positive
55868	F	35	15674	Positive
44551	M	45	15623	Negative
58524	M	45	15623	Negative
25584	F	61	15633	Negative
58744	F	61	15643	Positive
87524	M	19	15762	Positive
87384	M	19	15762	Negative
17583	F	19	15762	Positive

M: male; F: female

Fig. 2: Example of Tabular data

Qualitative data, also known as categorical data, provides answers to questions like "What/Which category?" Examples include abnormality levels in test results (e.g., normal, high, low) or distinguishing between cancer and normal tissue. Qualitative data can further be divided into nominal (no order between categories, e.g., hair color) and ordinal (ranked or ordered, e.g., blood pressure levels). Binary categories are a special case within qualitative data when there are only two options (e.g., positive or negative).

In many cases, tabular data files in healthcare contain a combination of these data subtypes, which are used to train machine learning algorithms for various tasks, such as predicting cancer or sepsis. The values are typically organized in rows and columns within a table. Each row represents an individual, and each column represents different variables (features/independent variables or target/dependent variable) for a given patient. The features in the dataset are mapped to the target of interest through a function to create a machine learning model capable of making predictions on new, unseen data. In supervised machine learning models, this relationship can be represented as: $Y = f(X) + e$, where Y is the target of interest, X is the features, f is the function representing the mathematical relationship between X and Y, and e represents the irreducible error inherent in any model. Before applying machine learning algorithms and training steps to map features to the target, data pre-processing tasks, including cleaning and standardization, are typically performed.

2- Data Cleaning

The cleaning tasks for numerical data encompass various activities aimed at optimizing the data for machine learning training in tabular studies. These tasks include reducing data noise through feature selector tools, addressing missing data through techniques like imputation or removal, scaling the data, converting text to numerical representations, and statistically assessing features to minimize issues like multicollinearity. By employing these approaches collectively, the ML training process can be optimized, leading to the development of a more generalizable model.

Traditionally, many of these tasks required manual approaches, which could be time-consuming. However, with the emergence of powerful data science applications, we now have the ability to expedite these tedious processes using automated, standardized, and validated approaches. These advanced applications offer user-friendly and scientifically robust toolkits specifically designed to address the aforementioned preprocessing tasks.

2.1 Missing values: Missing values are common in real-world datasets due to limitations in data capture or other factors. Handling missing values is crucial to make optimal use of available data [15]. Here are some approaches:

Drop samples with missing values: This is suitable when the number of missing values in a sample is high and doesn't significantly affect the dataset.

Replace missing values with zero: Applicable when zero represents the absence of a value, but it may not be suitable in all cases.

Replace missing values with mean, median, or mode: Statistical measures can be used as replacements, providing more meaningful approximations.

Interpolate missing values: Generate values within a range based on a step size, assuming a certain pattern or trend.

Extrapolate missing values: Populate values beyond a given range by referencing another variable or the target variable.

Use other features to predict missing values: Train a model using other variables to predict and approximate missing values accurately.

2.2 Noisy data: Noisy data refers to random errors or variances in measured variables. Some techniques [16] to handle noisy data include:

Binning: Divide data into equal-sized bins and replace values within each bin with mean, median, or boundary values.

Regression: Fit data points to a regression function to smoothen noise, using linear or polynomial equations depending on the variables involved.

Clustering: Create groups or clusters of similar data points, identifying outliers or noisy data points outside these clusters.

2.3 Removing outliers: Outliers are data points that deviate significantly from the expected patterns in the dataset. They can disrupt predictions and calculations. Techniques to handle outliers include:

Clustering: Group data points with similar values, treating points outside the clusters as outliers.

Box plots: Use box plots to identify outliers based on median, interquartile ranges, and extreme values. Remove outliers by filtering the variable based on the maximum and minimum range.

By employing these techniques, data cleaning aims to enhance data quality, ensuring accurate and reliable analysis for machine learning and other applications.

3- Data Transformation

After completing the data cleaning process, it is necessary to transform the high-quality data into alternative forms by changing the values, structure, or format of the data. The following strategies are commonly used for data transformation:

3.1 **Generalization**: By utilizing concept hierarchies, we can convert low-level or granular data into higher-level information. For example, transforming an address attribute from a city level to a country level.

3.2 **Normalization**: Normalization is a crucial data transformation technique widely employed. It involves scaling numerical attributes to fit within a specified range. The purpose is to establish correlations between different data points by constraining the attribute values. Various normalization methods include:

- Min-max normalization
- Z-Score normalization
- Decimal scaling normalization

3.3 **Attribute Selection**: New data properties are derived from existing attributes to facilitate the data mining process. For instance, transforming the date of birth attribute into a property like "is_senior_citizen" for each tuple, which can directly impact predictions related to diseases or survival chances.

3.4 **Aggregation**: Aggregation involves storing and presenting data in a summarized format. For example, sales data can be aggregated and transformed to display information on a monthly or yearly basis.

3.5 **Data Reduction**: In data warehousing, datasets can be extremely large and challenging to handle with data analysis and mining algorithms. Data reduction techniques are employed to obtain a smaller representation of the dataset without compromising the quality of analytical results. Some data reduction strategies include:

- Data cube aggregation: Summarizing data in a compact form.
- Dimensionality reduction: Reducing the number of features or attributes using techniques such as Principal Component Analysis.
- Data compression: Reducing data size through encoding technologies (lossless or lossy compression).
- Discretization: Converting continuous attributes into intervals for easier interpretation and analysis.
- Numerosity reduction: Representing data using models or equations instead of storing the entire dataset.
- Attribute subset selection: Selecting only relevant attributes for model training, avoiding high-dimensional data and potential underfitting or overfitting issues.

By employing these data transformation strategies, we can effectively prepare the data for analysis and enhance the accuracy and efficiency of machine learning algorithms.

4- Balanced or imbalanced data

Imbalanced data refers to a situation where categorical fields exhibit an uneven distribution of observations across different classes. This imbalance can pose significant challenges for models and analyses. In such cases, models can become lazy and achieve good performance simply by predicting the majority class as a default. To illustrate, imagine a dataset where 90% of the observations belong to one target class and only 10% to the other. Even if we consistently predict the majority class, we would still achieve 90% accuracy. This demonstrates how a model may perform well without utilizing any meaningful information from the features.

Class imbalance can also affect the learning of features. Models rely on identifying patterns, and when classes are severely underrepresented, it becomes challenging for the models to make accurate predictions for these minority groups. This challenge becomes more pronounced when multiple features exhibit imbalanced distributions, potentially resulting in situations where rare combinations of classes occur only in a few observations.

There are various techniques available to address imbalanced data. Under sampling involves reducing the number of observations in the overrepresented classes to achieve a more balanced distribution. On the other hand, oversampling entails generating additional data for the underrepresented classes. Several approaches can be employed to achieve this, such as leveraging Python packages like imbalanced-learn or utilizing services like Gretel. Imbalanced features can also be rectified through feature engineering, which aims to combine classes within a field without losing valuable information.

By applying appropriate sampling techniques and feature engineering strategies, imbalanced data can be effectively mitigated, leading to more accurate and reliable model predictions.

III. Bias and Ethics in Machine Learning

Machine learning algorithms are often perceived as objective and unbiased, but this assumption is incorrect. In healthcare, existing biases within healthcare systems, influenced by social determinants of health and protected characteristics related to human rights, can be embedded within these algorithms. Factors such as race, ethnicity, gender, socioeconomic status, access to healthcare, and

attitudes towards healthcare can introduce inherent biases. Developers must understand the composition of their training data and the clinical context to avoid making erroneous approximations, particularly for underrepresented populations.

Ethical considerations also come into play when developing machine learning models. Developers need to evaluate the potential impact of their models on both physicians and patients. What are the ethical implications if harmful outcomes arise from the predictions made by the model? By addressing these concerns, developers can establish appropriate clinical parameters that indicate when and where the model is suitable for use. Following these parameters and implementing proper oversight can help ensure ethical outcomes align with expected clinical outputs.

However, there is a risk that machine learning tools, if not carefully designed, may detect and amplify existing biases within the healthcare system. If error rates vary across different subpopulations, it becomes crucial to consider the interpretation and implications of the model. Numerous instances, both within and outside of healthcare, have demonstrated bias in machine learning algorithms. Examples include algorithms used for predicting recidivism in the justice system inaccurately predicting re-offense for non-white offenders, problematic outcomes for non-white subsets of the population in the Framingham Heart Study predictions, and gender biases in Google algorithms towards certain occupations.

Considering the social determinants of health mentioned earlier, it is vital to reflect on the future implications of biases in machine learning models. In jurisdictions without government healthcare insurance, there may be incentives for some parties to intentionally design algorithms that prioritize profit over patient well-being. Racial minorities or individuals living in rural areas may be underrepresented in training data, leading to incorrect predictions, diagnoses, or treatment recommendations.

To address these concerns, it is crucial to continuously evaluate and mitigate biases in machine learning models. This requires robust data collection, diverse representation in training data, ongoing monitoring, and accountability. By striving for fairness, transparency, and ethical responsibility, machine learning algorithms can be developed to promote equitable healthcare outcomes for all.

IV. Future Directions and Conclusion

Data preprocessing forms the foundation of successful machine learning applications in healthcare. As emerging technologies and trends continue to shape this field, it is crucial to adapt and harness their potential. By integrating federated learning, explainable AI, and real-time data streams, healthcare analytics can be revolutionized, leading to improved patient outcomes and personalized medicine. However, these advancements must go hand in hand with ethical considerations. Upholding data privacy, mitigating biases, and ensuring responsible data usage are essential to build trust and achieve the full potential of machine learning in healthcare. Continuous research, collaboration, and adherence to ethical guidelines will drive the future of data preprocessing in healthcare, paving the way for a transformative impact on the field and, ultimately, on patient well-being.

References

[1] S.V. Pillai, R.S. Kumar The role of data-driven artificial intelligence on COVID-19 disease management in public sphere: a review Decision, 48 (2021), pp. 375-389, 10.1007/S40622-021-00289-3

[2] E. Mbunge, J. Batani, G. Musuka, I. Chitungo, I. Chingombe, B.M Tafadzwa Dzinamarira, D. Lamprou Emerging technologies for tackling pandemics Emerging Drug Delivery and Biomedical Engineering Technologies: Transforming Therapy, CRC Press (2023), pp. 211-219

[3] J. Batani, M.S. Maharaj Towards data-driven models for diverging emerging technologies for maternal, neonatal and child health services in sub-Saharan Africa: a systematic review Glob. Health J. (2022), 10.1016/J.GLOHJ.2022.11.003

[4] M. I. Jordan T. M. Mitchell, Machine learning: Trends, perspectives, and prospects.Science349,255-260(2015).DOI:10.1126/science.aaa8415

[5] Dash, S., Shakyawar, S.K., Sharma, M. *et al.* Big data in healthcare: management, analysis and future prospects. *J Big Data* **6**, 54 (2019). https://doi.org/10.1186/s40537-019-0217-0

[6] Mohd Javaid, Abid Haleem, Ravi Pratap Singh, Rajiv Suman, Shanay Rab, Significance of machine learning in healthcare: Features, pillars and applications, International Journal of Intelligent Networks, Volume 3, 2022, Pages 58-73, ISSN 2666-6030, https://doi.org/10.1016/j.ijin.2022.05.002.

[7] Ahsan MM, Luna SA, Siddique Z. Machine-Learning-Based Disease Diagnosis: A Comprehensive Review. Healthcare (Basel). 2022 Mar 15;10(3):541. doi: 10.3390/healthcare10030541. PMID: 35327018; PMCID: PMC8950225.

[8] Janiesch, C., Zschech, P. & Heinrich, K. Machine learning and deep learning. *Electron Markets* **31**, 685–695 (2021). https://doi.org/10.1007/s12525-021-00475-2

[9] Van den Broeck J, Cunningham SA, Eeckels R, Herbst K. Data cleaning: detecting, diagnosing, and editing data abnormalities. PLoS Med. 2005 Oct;2(10):e267. doi: 10.1371/journal.pmed.0020267. Epub 2005 Sep 6. PMID: 16138788; PMCID: PMC1198040.

[10] Emmanuel, T., Maupong, T., Mpoeleng, D. *et al.* A survey on missing data in machine learning. *J Big Data* **8**, 140 (2021). https://doi.org/10.1186/s40537-021-00516-9

[11] Prakash M Nadkarni and others, Natural language processing: an introduction, *Journal of the American Medical Informatics Association*, Volume 18, Issue 5, September 2011, Pages 544–551, https://doi.org/10.1136/amiajnl-2011-000464

[12] Yim W, Yetisgen M, Harris WP, Kwan SW. Natural Language Processing in Oncology: A Review. *JAMA Oncol.* 2016;2(6):797–804. doi:10.1001/jamaoncol.2016.0213

[13] Massimo Salvi, U. Rajendra Acharya, Filippo Molinari, Kristen M. Meiburger, The impact of pre- and post-image processing techniques on deep learning frameworks: A comprehensive review for digital pathology image analysis, Computers in Biology and Medicine, Volume 128, 2021, 104129, ISSN 0010-4825, https://doi.org/10.1016/j.compbiomed.2020.104129.

[14] Giger ML, Chan HP, Boone J. Anniversary paper: History and status of CAD and quantitative image analysis: the role of Medical Physics and AAPM. Med Phys. 2008 Dec;35(12):5799-820. doi: 10.1118/1.3013555. PMID: 19175137; PMCID: PMC2673617

[15] Emmanuel, T., Maupong, T., Mpoeleng, D. *et al.* A survey on missing data in machine learning. *J Big Data* **8**, 140 (2021). https://doi.org/10.1186/s40537-021-00516-9

[16] Kiran Maharana, Surajit Mondal, Bhushankumar Nemade, A review: Data pre-processing and data augmentation techniques, Global Transitions Proceedings, Volume 3, Issue 1, 2022, Pages 91-99, ISSN 2666-285X, https://doi.org/10.1016/j.gltp.2022.04.020.