Data Mining and Processing in the Age of Big Data and Artificial Intelligence – Issues, Privacy, and Ethical Considerations

Conference Paper · August 2024

DOI: 10.1109/ASIANCON62057.2024.10838087

CITATIONS

0

3 authors:

Saransh Arora

Amazon Web Services

11 PUBLICATIONS **45** CITATIONS

SEE PROFILE

Sunil Raj Thota

Amazon

11 PUBLICATIONS **48** CITATIONS

SEE PROFILE

Sandeep Gupta

Samrat Ashok Technological Institute

30 PUBLICATIONS **243** CITATIONS

SEE PROFILE

2024 4th Asian Conference on Innovation in Technology (ASIANCON)
Pune, India. Aug 23 - 25, 2024

Data Mining and Processing in the Age of Big Data and Artificial Intelligence - Issues, Privacy, and Ethical Considerations

Saransh Arora
Jaypee institute of information technology
Noida, U.P. India
Saransha.1994@gmail.com

Sunil Raj Thota
Technology, Engineering
Andhra university
Visakhapatnam, AP, India
thotasunilraj@gmail.com

Sandeep Gupta
Techieshubhdeep it Solutions Pvt. Ltd,
Gwalior, M.P. India
ceo.techies@gmail.com

2024 4th Asian Conference on Innovation in Technology (ASIANCON) | 979-8-3503-5421-8/24/$31.00 ©2024 IEEE | DOI: 10.1109/ASIANCON62057.2024.10838087

***Abstract*—The integration of big data and AI, which is at present a most crucial development in data mining and processing, can be regarded as the beginning of a new era in the field of science and technology. The research primarily focuses on using ML and data mining methods to forecast cardiac diseases. It also identifies the most efficient and effective approaches for managing enormous datasets and producing accurate forecasts. The results show that LightGBM comes in second with 93.8% accuracy and AdaBoost with 95.2% accuracy. The accuracy rates of other models, such KNN and Naive Bayes, are 88.52% and 90.56%, respectively. These results highlight how sophisticated machine learning algorithms might improve a predicted accuracy of heart disease diagnosis. The paper underscores the necessity of safeguarding data and the importance of stakeholder collaboration to ensure that the advancements in big data and AI are equitable, sustainable, and centred around human well-being.**

Keywords—Big Data, artificial intelligence (AI), DM, data processing, Issues, Data Privacy, and Ethical Considerations.

I. INTRODUCTION

Big data and AI bring revolutionary promise regarding data mining and processing, yet their integration poses another set of problems. As data mining and processing activities are increasing across all sectors, the urgent is to develop privacy and ethical rules. Among the crucial components in data mining and processing are informed consent, data anonymisation, algorithmic bias, data ownership, and regulatory compliance. We have to be considered to make certain that the activities of data mining and processing do not violate human rights, promote transparency, and prevent possible harm. The development of ethical values will be important for the generation of trust, honesty and conscious development in an era of big data and AI. Machine learning as part of AI. The healthcare business generates a vast amount of data, and ML has shown to be useful in helping to make predictions and judgements from this data. An employ of data mining and ML methods to forecast a risk of contracting certain diseases has gained popularity recently. Critical as their work is as doctors diagnosing and prognosing cardiovascular disease, cardiologists may do a better job of treating patients if they accurately categorise them. In order to forecast heart disease, a variety of techniques have been utilised for knowledge abstraction, utilising well-established data mining approaches. This effort aims to discover the best and most efficient approach or tactics for predicting heart attacks using techniques like AI, data mining (DM), preprocessing, and large illness datasets.

A. Aim and Scope of the Paper

A motivation behind this study is to address a burgeoning challenges and opportunities presented by the integration of big data and AI technologies in the field of data mining, specifically within healthcare. As the healthcare industry generates vast amounts of data, there is a critical need for effective tools that can predict, diagnose, and manage diseases like cardiovascular conditions. This study is important because it has the ability to improve decision-making and predictive analytics, which in turn might improve patient outcomes and make better use of healthcare resources. The paper aims to offers an overview and assessment of a state-of-the-art research on heart disease prediction utilising sophisticated data mining and ML methods. It also aims to identify the most efficient and successful approaches for managing enormous datasets and producing precise forecasts. The highlighting the contributions of this paper:

- The paper includes a comprehensive approach of how to forecast heart disease utilising data mining methods and ML algorithms. Indeed, it includes data preprocessing, feature selection using RF, and use of more complex algorithms such as AdaBoost, LightGBM. This systematic approach is relevant for skills development and can be improved upon by other researchers who wish to develop a similar approach for analysing and modelling big data.
- The paper highlights the necessity of considering the ethical and privacy issues when it comes to data mining and processing with reference to the rather sensitive aspect of health data. Some of them include consent, obscuring of data, declarative specification of the algorithm, and rules for using legal norms.
- The review of literature section presents a synthesis of current state-of-the-art research on topic covering ethical considerations and concerns related to data collection, and analysis. It summarises different researches to provide coherent information and overview on the topic of ethics of big data and AI.
- The integration of boosting algorithms such as AdaBoost and LightGBM in predicting the heart disease is the creativity of the contemporary machine learning methods. By demonstrating the effectiveness of these algorithms in handling large datasets and providing high accuracy, the paper paves the way for their broader adoption in healthcare analytics.

These contributions collectively advance the field of data mining and machine learning in healthcare, promoting both technical innovation and ethical responsibility.

B. Organization of paper

A following paper organised as: Section II and III provide the overview of data mining and data preprocessing technique with their issues, data privacy, and ethical considerations, then Section IV and V discuss the methodology and results, Next Section VI provide the some existing works on related research areas , at last section VII provide the conclusion and future work.

II. Overview Of Data Mining And Processing

The process of extracting valuable patterns and models from an enormous dataset is known as data mining. Data mining is fundamentally contingent on data quality. Raw data is frequently contaminated with outlier data, missing values, noise, incompleteness, and inconsistency. Therefore, it is critical that these data be processed prior to mining [1]. Data mining and processing form the core of the data analytics pipeline, in efforts to extract discover hidden gems of meaning and eventually, knowledge from huge and complex datasets. Data mining and processing are of inevitable importance to any company in all industries as an efficient tool to gain a competitive advantage, come up with new ideas and accurately make informed decisions in the age of digital transformation with its overflow of data[2].

A. Data Mining

DM refers to the process of obtaining such important information from large datasets. A number of terms describe this procedure; some examples include data/pattern analysis, information discovery, knowledge extraction, and knowledge mining from data (Figure 1) [3].

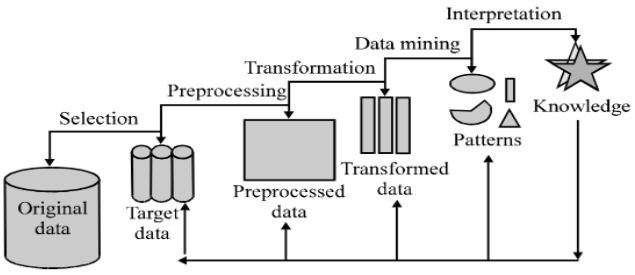

Fig. 1. Knowledge discovery Process[4]

Data mining is a systematic procedure utilised to extract valuable information from vast quantities of data. The objective of this methodology is to identify patterns that were hitherto undiscovered. After identification of these patterns, we can subsequently be utilised to inform strategic business development decisions. There are three stages involved:

- **Exploration:** Data is cleansed and transformed into a different format as the initial stage of data exploration. Subsequently, critical variables and the nature of the data in relation to problem at hand are ascertained.
- **Pattern Identification:** The identification of patterns occurs as the second stage, following the exploration, refinement, and definition of data for a specific variable. Determine which predictive patterns produce the most precise outcomes and adopt them.
- **Deployment:** Patterns are deployed for desired outcome.

The utilisation of DM in conjunction with big data has become prevalent throughout entire lifecycle of electronic products, encompassing design, production, and service[5].

B. Data Preprocessing

Data preprocessing consists of a straightforward conversion of unstructured data into a format that is comprehensible. The data preprocessing procedures are illustrated in the diagram presented in Figure 2.

Fig. 2. Data Preprocessing Steps[6].

It is critical to perform data preprocessing in order to increase data efficacy. A crucial stage in the process of data mining, data preprocessing entails transforming and preparing the dataset in an effort to increase the efficiency of knowledge discovery. Preprocessing comprises of a number of jobs such as reduction, cleansing, integration and transformation.

Data cleaning: Data cleansing is that process which deals in among other things with fixing incorrect records, defective pieces of information and then removing them from the database table or set of records.

Data Integration: A single data set that provides a unified perspective on the data is the main objective of data integration, which is the consolidation of information from various sources.

Data transformation: Data transformation is a critical step that takes the static raw data and translates it into a format that is human readable.

Data reduction: Data reduction is the procedure by which digital information is systematically organised and simplified. The acquisition of this data typically occurs via empirical and experimental methods.

Data discretisation: When a substantial volume of numeric data must be categorised solely on basis of nominal values, data discretisation becomes a critical concept. The nominal value signifies values of the discrete sets from which continuous data is subtracted in this scenario.

III. Methodology

The methodology of data mining and processing in the Big Data and AI context includes several steps necessary to extract the potentially useful information and construct the elements of the heart disease prediction model. This systematic approach makes it possible to process and manage large and complicated data and build very efficient AI solutions in the tasks of prediction. These process shows in figure 3 also discuss below:

A. Description of Dataset and Visualisation

There is a wealth of information available in the heart disease dataset about lifestyle choices and cardiovascular health. This includes details about individual patients like their gender, age, blood pressure, cholesterol, and heart rate as well as variables like family history, diabetes, obesity, smoking, and alcohol intake. Factors related to one's way of life are also considered, including the amount of time spent exercising, stress levels, eating habits, and amount of time spent sitting still.. In all, 8763 patient records from all corners of the globe make up the collection. The presence or absence of a risk of heart attack is indicated by a crucial binary classification characteristic that is included. Research and predictive analysis pertaining to cardiovascular health may greatly benefit from this dataset.

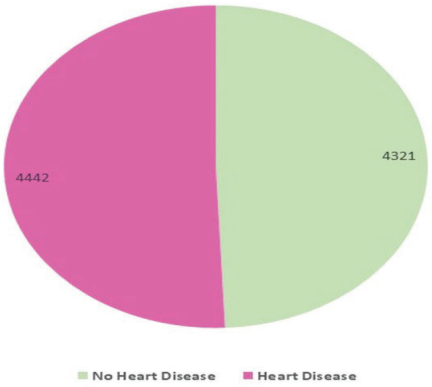

Fig. 3. Pie chart of data distribution of classes in the heart disease dataset.

The dataset for this experiment has a decent mix of classes, as illustrated in Figure 4, where class 1 represents heart diseases (4442 cases) and class 0 represents no heart disease (4321 instances).

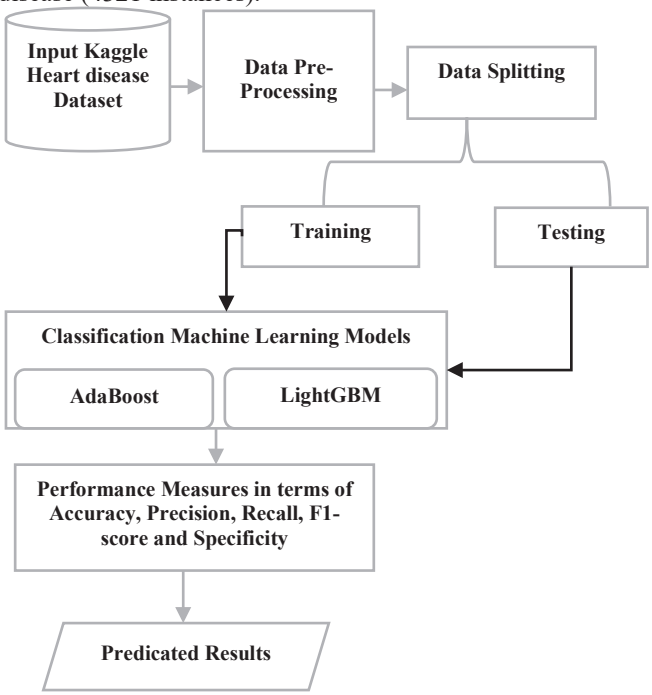

Fig. 4. Flowchart of heart disease prediction with ML techniques

B. Data Preprocessing

The preprocessing step's objective is to determine what data the automobile needs before deciding whether or not to utilise it. For preprocessing, data must be properly cleaned and prepared. Following are some of the stages involved in data preprocessing:

- **Remove null values ():** When certain fields in a data record are blank, null values will show. Data may have null values due to a number of factors, including mistakes in data input or missing data. These null values can skew analysis and lead to inaccurate results. By removing them, can ensure that analysis is based on reliable data.
- **Label encoder ():**Label encoding is a crucial part of the data pre-processing procedure for supervised learning models. The utility class label encoder may be used to transform labels into integers between zero and one.
- **Dataset Splitting:** The ratio of dataset splitting has been carefully selected, considering factors like overfitting, model complexity, dataset size, and the particular needs of ML tasks. A 60:40 ratio is used to partition the dataset. This implies that 40% of the dataset is utilised for method testing and 60% of the dataset is utilised for training each algorithm.

C. Classification machine learning models

Two boosting methods based on ML were investigated in this research for an aim of heart disease prediction. Below are the algorithms that were used in the experiment.

1) AdaBoost

One ensemble approach that uses ML methods is AdaBoost [7] often known as adaptive boosting. The majority of AdaBoost estimators are one-level decision trees, which are also called decision trees with a single split. Decision stumps is a common name for these trees. This method constructs a model by treating all data points as equally important. Then, more weight is given to the points that were wrongly allocated. All points with larger weights are given a larger weight in the following model. Until the stated error is decreased, it won't cease training models [8].

2) Light GBM

The best way to maximise Light GBM's performance is via distributed systems [9]. Decision trees may be trained to expand "leafwise" using Light GBM, which means that the tree is only divided once depending on the gain for each given situation. Leaf-wise trees are susceptible to overfitting, particularly when dealing with smaller datasets. Tree depth reduction may help avoid overfitting. By splitting data into bins according to the distribution's histogram, Light GBM uses a histogram-based technique. The data splitting, gain computation, and iteration processes employ the bins rather than each individual data point.

D. Model Evaluation

A crucial step in ML, model assessment is all about seeing how accurate outcomes are predicted by well-trained models. Models' ability to generalise to new data effectively is ensured at this critical step, which guides decisions about deployment and improvement.

IV. Results and Discussion

Provide a findings and analysis of a heart disease prediction utilising data mining approaches in this part, taking into account measurements of f1-score, sensitivity, specificity, accuracy, and precision. The following is a discussion of the contributing elements to these:

A. Performance Evaluation

A set of formulas called performance evaluation is used to calculate how successful a classifier or model is [42]. The definitions of certain key concepts that are utilised in the performance assessment equations are provided below:

- **True Positive (TP):** The individual is both expected to be and is in good health.
- **False Positive (FP):** The individual is well, but they are predicted to get sick.
- **True Negative (TN):** A person is unwell and is expected to be sick.
- **False Negative (FN):** A person is unwell, but expected to recover soon.

Confusion Matrix

Classifier and model performance in classifying datasets may be evaluated using the confusion matrix. TN and TP are sent to the right category, whereas FN and FP are directed to the incorrect category. For the accurate classifier or model, TP and TN are classified more than FN and FP[10][11], as shown in Table 1.

TABLE I. CONFUSION MATRIX

Classes	Negative (Actual)	Positive (Actual)
Negative (Predict)	TN	FN
Positive (Predict)	FP	TP

Accuracy: The amount of correct predictions made by the model or classifier is measured by its accuracy. Applying (1) allows one to determine the accuracy.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \dots\dots\dots\dots (1)$$

Precision: The degree to which a diagnosis or expected outcome is near to the actual outcome may be determined by measuring its precision. Use the formula (2) to determine the precision.

$$Precision = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \dots\dots\dots\dots\dots\dots\dots\dots (2)$$

F-Measure: The F-Measure is the average degree of agreement between the two metrics, Precision and Recall. Applying (3) allows one to compute the F-measure.

$$F1 - Score = \frac{2(Precision * Recall)}{\mathrm{Precision} + \mathrm{Recall}} \dots\dots\dots\dots (3)$$

Sensitivity(Recall): A true positive rate indicator is sensitivity. Applying (4) yields the sensitivity value.

$$Recall = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \dots\dots\dots\dots\dots\dots\dots\dots (4)$$

Specificity: A true negative rate indicator is specificity. Put otherwise, the frequency with which a person is diagnosed or predicted to be ill. Applying (5) yields the specificity.

$$Specificity = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}} \dots\dots\dots\dots\dots\dots\dots\dots (5)$$

These performance measures evaluate the performance of data mining models that shows in below table 2.

TABLE II. PARAMETER COMPARISON WITH LGBM AND ADABOOST CLASSIFIERS OF HEART DISEASE

Parameters	Light GBM	AdaBoost
Accuracy	93.8	95.2
Sensitivity	93.4	95.2
Precision	98.7	98.7
Specificity	95.3	95.3
FI-Score	96	96.98

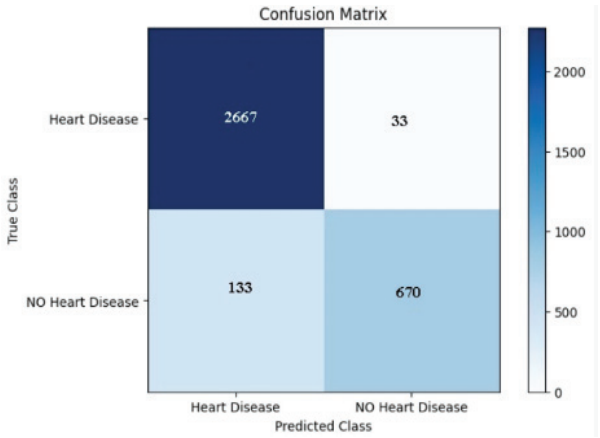

Fig. 5. Confusion matrix of LGBM classifier for heart disease

The confusion matrix for the LGBM classifier in predicting heart disease shows 2667 TP, 33 FN, 133 false positives (FP), and 670 TN, shows in figure 5. It reveals how well the model distinguishes between individuals with and without heart disease, with a focus on correctly identifying cases and minimising misclassifications.

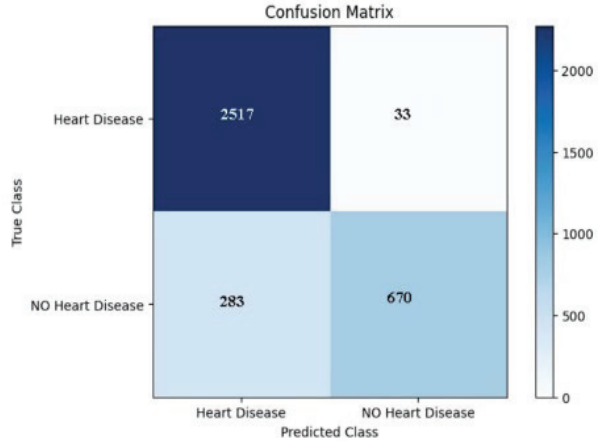

Fig. 6. Confusion matrix of AdaBoost classifiers for heart disease

The confusion matrix for the AdaBoost classifier in predicting HD illustrates its performance across four key metrics, shows in figure 6. It correctly identifies 2517 TP, accurately predicting individuals have HD. However, it also shows 33 FN, indicating cases where individuals with HD were incorrectly classified as not having it. There are 283 FP, where individuals without HD were incorrectly predicted to have it. The matrix also includes 670 TN, representing individuals correctly identified as not having HD.

The accuracy comparison of many data mining methods for heart disease prediction is shown in Table 3. The AdaBoost model achieves the highest accuracy at 95.2%, followed closely by Light GBM at 93.8%. Naive Bayes performs well with an accuracy of 90.56%, while the

TABLE III. ACCURACY COMPARISON WITH DIFFERENT DM MODELS OF HEART DISEASE

Models	Accuracy (%)
Light GBM	93.8
AdaBoost	95.2
KNN[12]	88.52
Support vector Machine [13]	83
Naive Bayes [14]	90.56
Vote[15]	87.4
Decision Trees[16]	86.37
HRFLM[17]	88.7

HRFLM model records 88.7%. The KNN model has an accuracy of 88.52%, slightly above the Vote model at 87.4% and Decision Trees at 86.37%. The Support Vector Machine shows the lowest accuracy among the compared models at 83%. Figure 8 illustrates these accuracies in a bar graph, highlighting the performance differences across the models.

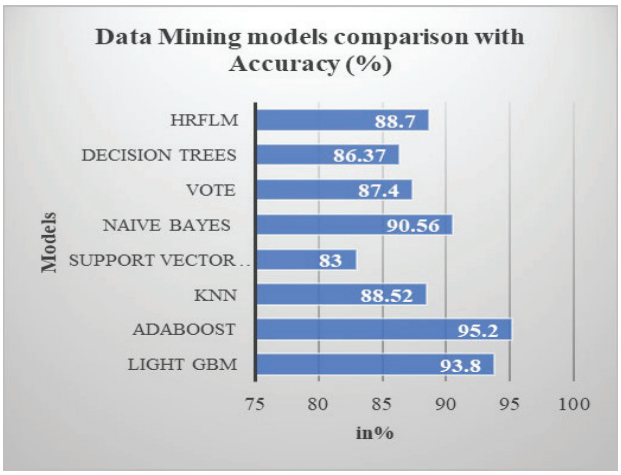

Fig. 7. Bar graph of DM models compared with heart disease dataset

V. LITERATURE REVIEW

This section encompasses a literature review, focusing on data mining and processing in an age of big data and AI - issues, privacy, and ethical considerations across different regions.

In, Gold Nmesoma Okorie et al., (2024) to research and summarise the issues with contemporary data collecting and analysis as well as ethical procedures. This study presents a comprehensive literature review that delves into fundamental ethical concepts related to data collection and analysis. It covers topics such as the significance of permission and privacy, the challenges posed by big data, and the variations in ethical frameworks across various places[18].

In, Cary, Wen and Mahatanankoon, (2003) determining the potential dangers to a company that is suspected of conducting unethical business practices and the ethical issues associated with data mining. Incorporating ten data mining systems development practices into the software development lifecycle would prevent the materialisation of these risks, as suggested in the paper, which also examined pertinent ethical policies[19].

In, R. Surendiran et al. (2023), seeks to predict cardiac or cardiovascular disease utilising three AI tools: neural networks, DT, and NB. A number of particular factors are used to evaluate the determination of various techniques, and adjustments are made to improve accuracy. Thereafter, a comparison will be made between each approach's accuracy based on several attributes. The most accurate technique then ascertains whether an individual decides to acquire coronary heart disease or not. In the event that the patient persists, medical practitioners may use this technique to detect diseases early and provide timely treatment[20].

In, Okomayin and Kolade, (2023) this paper examines the legal, privacy, and ethical concerns surrounding data mining, evaluates the processes involved in data mining, and proposes solutions to the existing ethical and privacy issues in the field. The exclusive objective of this manuscript is to present a novel approach to data mining that imposes limitations on scientists in terms of ethical considerations, privacy, and legality[21].

In Osamah Sami et al. (2021), utilising data preprocessing approaches to improve an accuracy of ML methods for a prediction of coronary heart disease. As an illustration, the DT classifier improved the predictive accuracy of coronary heart disease by 1.39% compared to the previous work, the RF classifier by 2.7%, the KNN classifier by 2.58%, the MLP classifier by 2.64%, and the NB classifier by 0.66%[14].

Therefore, Koo, Kang and Kim, (2020) an analysis of pertinent research and validation of contemporary benchmarks set by international standardisation organizations enable the detection of security risks and hazards that manifest during the entire life cycle of big data. In addition, they designate five phases—collection, storage, analytics, utilisation, and annihilation—and classify them within a security taxonomy that corresponds to the identified threats and security concerns throughout the big data life cycle[22].

In, Li and Zhang, (2017) discuss the possible dangers and concerns associated with artificial intelligence applications while bringing up security, privacy, and ethnic problems. They outline our expectations for the evolution of artificial intelligence and propose countermeasures in research, policy, and monitoring[23].

In, Zhao, Gong and Wang, (2021) examines the current state of the AI sector. It then examines the many risks brought on by the privacy crisis by illustrating the issues with data leaking in the modern artificial intelligence era. In conclusion, this article outlines the legal and non-legal avenues for safeguarding the right to privacy, as well as the evolution of these avenues[24].

The cited research articles highlight several critical challenges and research gaps in the fields of data science, DM, big data, and AI. Some of overarching challenges and gaps include:

Ethical Practices and Frameworks: Although ethical aspect of data collection, analysis, AI applications, as well as ethical frameworks and standards are now widely recognised, there is a part of ethical issues that are still lacking in development and implementation.

Ethical Issues in Data Mining and AI: Identifying and resolving AI's and data mining's ethical imbroglios calls for sustainable and ethical conduct. These problems might be the examples of data privacy issues, algorithm bias and the possibility to restrict personal freedom and rights.

Legal and Regulatory Compliance: Data security and privacy issues in connection with legal requirements pose quite a challenge. Proper legal due diligence and standards

regulating innovations aiming at bringing about a balance while at the same time protecting the rights of individuals and the integrity of data are equally vital.

Security and Privacy Concerns: These issues of data security and privacy will always be the central topics when the AI applications are being more complex or wider data collecting happens.

Data Lifecycle Management: Data governance, quality assurance, and security like the entire lifecycle of data from the collection to disposal are the core issues relating to the managing of big data.

Ethnic and Societal Implications of AI: The cooperation of different disciplines is crucial in the complete understanding of all the social issues arising from AI that range from discrimination, inequality, morals of AI design, and development.

Data Leakage and Privacy Crisis: The AI industry is still seeing data leaks that clearly demonstrate how respecting individuals' privacy rights is crucial and should be done with secure methods.

In the big data and AI age, addressing these issues and research gaps calls for multidisciplinary cooperation, creative methodology, and a comprehensive approach to data governance, ethics, and legislation. To maximise the positive social effects of data-driven technology and to encourage ethical and responsible data practices, further study in these areas is necessary.

VI. Conclusion And Future Work

The article concludes that although the DM and processing will be extremely impactful in the age of big data and AI, there are some ethical and practical problems. They do not only look into the way data, AI, and DM function to find out whether or not these technologies push the boundaries of innovation, help in decision-making, and redefine whole industries. Data privacy, data security, and algorithm bias and legal requirements are the responsibilities that should be fulfilled to guarantee that data mining and processing are ethical. In the case of heart disease prediction, the authors show that using more elaborate ML algorithms is possible to accurately predict heart diseases with a mean accuracy level of 95. 2% using the AdaBoost algorithm. This high level of accuracy is evidence of the capability of the DM and AI methods to deliver rich insights and enhance decisions at critical sectors like the healthcare industry. Furthermore, there is the need to involve key players in the industry, policy makers, scientist and ethicist in putting up a detailed set of standards and framework that are applicable. It is crucial to form this partnership to manage the development and impact of techniques connected with big data and AI and to encourage innovative stages of data mining and information processing. Thus, the setting of these standards will help the field gain grounds in efforts to provide for fair, efficient, and appropriate applications of data mining and AI technologies.

References

[1] S. A. Alasadi and W. S. Bhaya, "Review of data preprocessing techniques in data mining," *J. Eng. Appl. Sci.*, 2017, doi: 10.3923/jeasci.2017.4102.4107.

[2] P. Swisstac Bravin, "A review on Data Preprocessing Techniques in Data Mining," *Int. J. Adv. Eng. Manag.*, 2022.

[3] E. Dehaerne, B. Dey, S. Halder, S. De Gendt, and W. Meert, "Code Generation Using Machine Learning: A Systematic Review," *IEEE Access*. 2022. doi: 10.1109/ACCESS.2022.3196347.

[4] N. Rahman, "Data Mining Techniques and Applications," *Int. J. Strateg. Inf. Technol. Appl.*, 2018, doi: 10.4018/ijsita.2018010104.

[5] S. Lv, H. Kim, B. Zheng, and H. Jin, "A review of data mining with Big Data towards its applications in the electronics industry," *Applied Sciences (Switzerland)*. 2018. doi: 10.3390/app8040582.

[6] V. Agarwal, "Research on Data Preprocessing and Categorization Technique for Smartphone Review Analysis," *Int. J. Comput. Appl.*, 2015, doi: 10.5120/ijca2015907309.

[7] X. Sun and H. Zhou, "Experiments with Two New Boosting Algorithms," *Intell. Inf. Manag.*, 2010, doi: 10.4236/iim.2010.26047.

[8] S. M. Ganie, P. K. D. Pramanik, S. Mallik, and Z. Zhao, "Chronic kidney disease prediction using boosting techniques based on clinical parameters," *PLoS One*, 2023, doi: 10.1371/journal.pone.0295234.

[9] G. Ke *et al.*, "LightGBM: A highly efficient gradient boosting decision tree," in *Advances in Neural Information Processing Systems*, 2017.

[10] W. Chen, S. Chen, H. Zhang, and T. Wu, "A hybrid prediction model for type 2 diabetes using K-means and decision tree," in *Proceedings of the IEEE International Conference on Software Engineering and Service Sciences, ICSESS*, 2017. doi: 10.1109/ICSESS.2017.8342938.

[11] H. M and S. M.N, "A Review on Evaluation Metrics for Data Classification Evaluations," *Int. J. Data Min. Knowl. Manag. Process*, 2015, doi: 10.5121/ijdkp.2015.5201.

[12] S. Ouf and A. I. B. ElSeddawy, "A PROPOSED PARADIGM FOR INTELLIGENT HEART DISEASE PREDICTION SYSTEM USING DATA MINING TECHNIQUES," *J. Southwest Jiaotong Univ.*, 2021, doi: 10.35741/issn.0258-2724.56.4.19.

[13] K. Srivastava* and D. K. Choubey*, "Heart Disease Prediction using Machine Learning and Data Mining," *Int. J. Recent Technol. Eng.*, vol. 9, no. 1, pp. 212–219, 2020, doi: 10.35940/ijrte.f9199.059120.

[14] O. Sami, Y. Elsheikh, and F. Almasalha, "The Role of Data Pre-processing Techniques in Improving Machine Learning Accuracy for Predicting Coronary Heart Disease," *Int. J. Adv. Comput. Sci. Appl.*, 2021, doi: 10.14569/IJACSA.2021.0120695.

[15] M. S. Amin, Y. K. Chiam, and K. D. Varathan, "Identification of significant features and data mining techniques in predicting heart disease," *Telemat. Informatics*, 2019, doi: 10.1016/j.tele.2018.11.007.

[16] C. M. Bhatt, P. Patel, T. Ghetia, and P. L. Mazzeo, "Effective Heart Disease Prediction Using Machine Learning Techniques," *Algorithms*, 2023, doi: 10.3390/a16020088.

[17] S. Mohan, C. Thirumalai, and G. Srivastava, "Effective heart disease prediction using hybrid machine learning techniques," *IEEE Access*, 2019, doi: 10.1109/ACCESS.2019.2923707.

[18] Gold Nmesoma Okorie, Chioma Ann Udeh, Ejuma Martha Adaga, Obinna Donald DaraOjimba, and Osato Itohan Oriekhoe, "ETHICAL CONSIDERATIONS IN DATA COLLECTION AND ANALYSIS: A REVIEW: INVESTIGATING ETHICAL PRACTICES AND CHALLENGES IN MODERN DATA COLLECTION AND ANALYSIS," *Int. J. Appl. Res. Soc. Sci.*, 2024, doi: 10.51594/ijarss.v6i1.688.

[19] C. Cary, H. J. Wen, and P. Mahatanankoon, "Data mining: Consumer privacy, ethical policy, and systems development practices," *Hum. Syst. Manag.*, 2003, doi: 10.3233/hsm-2003-22402.

[20] S. R, "Estimating Heart Disease Used by Data Mining and Artificial Intelligence Techniques," *Int. J. Comput. Sci. Eng.*, vol. 10, no. 4, pp. 1–7, 2023, doi: 10.14445/23488387/ijcse-v10i4p101.

[21] A. Okomayin and A. Kolade, "Data Mining in the Context of Legality, Privacy, and Ethics," 2023, [Online]. Available: https://www.researchgate.net/publication/370844342

[22] J. Koo, G. Kang, and Y. G. Kim, "Security and privacy in big data life cycle: A survey and open challenges," *Sustainability (Switzerland)*. 2020. doi: 10.3390/su122410571.

[23] X. Li and T. Zhang, "An exploration on artificial intelligence application: From security, privacy and ethic perspective," in *2017 2nd IEEE International Conference on Cloud Computing and Big Data Analysis, ICCCBDA 2017*, 2017. doi: 10.1109/ICCCBDA.2017.7951949.

[24] Y. Zhao, C. Gong, and Y. Wang, "Privacy Crisis in the Age of Artificial Intelligence and its Countermeasures," in *Proceedings - 2021 7th Annual International Conference on Network and Information Systems for Computers, ICNISC 2021*, 2021. doi: 10.1109/ICNISC54316.2021.00035.