World Journal of Advanced Research and Reviews
eISSN: 2581-9615 CODEN (USA): WJARAI
Cross Ref DOI: 10.30574/wjarr
Journal homepage: https://wjarr.com/

(RESEARCH ARTICLE)

Check for updates

Machine learning models for personalised healthcare on marketable generative-AI with ethical implications

Kiran Gulia [1,*], Izhar Ahmad Hamdan [2], Nilay Datta [2], Yuvraj Gupta [2], Phool Kumar [3], Amita Yadav [2], Sandeep Mitten [4] and Rajiv Kumar [5]

[1] *Department of Engineering, Faculty of Engineering, Mathematics and Computer Science, University of Wolverhampton, UK.*
[2] *Department of Computer Science, Maharaja Surajmal Institute of Technology, New Delhi – India.*
[3] *Department of Marketing and Sales, University School of Management, University of Lille, France.*
[4] *Division of Clinical Research, Zydus Pharmaceutical Division of Clinical Research and Development, New Jersey, USA.*
[5] *Department of Sciences, University of Delhi, New Delhi, India.*

World Journal of Advanced Research and Reviews, 2024, 23(03), 707–720

Publication history: Received on 20 July 2024; revised on 30 August 2024; accepted on 02 September 2024

Article DOI: https://doi.org/10.30574/wjarr.2024.23.3.2660

Abstract

Personalised healthcare, underpinned by a deep understanding of individual patient variability, demands innovative solutions. Machine Learning (ML) models offer a promising avenue to achieve this by facilitating the development of enhanced Digital Twins (DTs). This research proposes a novel framework for creating DTs tailored to individual patients, considering not only physical attributes but also the intricate interplay of social, and biological factors. By capturing this comprehensive patient profile, ML-powered DTs have the potential to revolutionise healthcare by enabling predictive, preventive, and personalised care strategies, ultimately leading to improved patient outcomes and the development of a marketable, trust-worthy product. Current healthcare solutions lack personalisation, often failing to consider individual differences in disease presentation and response to treatment. Traditional DTs in healthcare often adopt a disease-centric or organ-specific approach, thereby restricting their capacity to deliver comprehensive, personalised care. To address this limitation, we propose a holistic Artificial Intelligence (AI) framework centered on ML models. Initially focusing on diabetes, our research aims to enhance diagnosis, treatment, and predictive capabilities through personalised insights, thereby optimising patient outcomes and care management. The benefits identified with our ML model are early disease prevention and risk stratification, optimised treatment planning and therapy selection, enhanced patient-physician communication and shared decision-making, reduced healthcare costs and improved resource allocation. Our models are designed to optimise patient care while prioritising safety and societal benefit. To ensure this, we have conducted a thorough assessment of potential ethical implications. Key challenges identified include data privacy and security, algorithmic bias, diagnostic accuracy, data interoperability and standardisation, integration with existing healthcare systems, ethical management of sensitive patient data, refinement of ML methodologies, addressing legal and ethical AI challenges, and suggesting robust ethical guidelines. A comprehensive evaluation of accuracy, reliability, and associated risks will be conducted prior to full-scale integration into the healthcare ecosystem to establish a robust ethical framework for our research models.

Keywords: Personal Digital Twins (PDT); Digital Twin (DT); Machine Learning (ML), Artificial Intelligence (AI), Base Diabetes Management using PDTs; Ethical implications.

* Corresponding author: Kiran Gulia

1. Introduction

Personalised Digital Twins (PDTs) in healthcare represent an advanced technological innovation designed to create virtual replicas of patients, simulating the human body with remarkable precision. These digital models are crucial for enhancing diagnostic accuracy and improving treatment efficacy by employing in silico simulations of disease progression and therapeutic interventions [1,2]. PDTs have emerged as invaluable tools not only in patient-specific care but also in medical research and development, offering a dynamic platform for exploring novel treatment strategies and accelerating drug discovery. Originating from the field of industrial design, the concept of Digital Twins (DTs) is now revolutionising healthcare by enabling the creation of comprehensive virtual representations of patients [3,4].

These PDTs integrate diverse data types, including genetic information, to provide personalised care tailored to each individual. For instance, recent advancements have led to the development of 3D heart models, which are employed to personalise cardiovascular care, allowing clinicians to simulate surgical procedures and predict outcomes with greater accuracy [5]. The creation of these digital replicas relies on sophisticated data security protocols that safeguard a patient's clinical history and sensitive information, ensuring the integrity and confidentiality of the data [6-8].

Research from Indiana University underscores the importance of integrating multi-scale data to construct accurate computational models that can predict how biological processes are influenced by various factors, including viral infections [9,10]. These digital models are then personalised with clinical data from individual patients, enabling the generation of precise predictions regarding diagnosis, prognosis, treatment efficacy, and the optimisation of therapeutic interventions. As a result, PDTs are positioned to significantly enhance personalised medicine, offering new avenues for patient care and medical research [11-15].

PDTs offer a transformative approach in healthcare, leveraging predictive algorithms to analyses real-time data, identify potential health risks, and prevent diseases before symptoms emerge. These advanced digital models can detect early signs of health deterioration, enabling timely interventions [16]. Moreover, PDTs, enriched with genetic information and detailed medical histories, allow for the personalisation of treatment plans, tailoring therapies to the unique needs of each patient. Individual differences in genetic makeup necessitate highly individualised care strategies, as standardised treatments can produce varying outcomes across different patients [17].

DTs offer clinicians the ability to virtually simulate various treatment options, assessing their effectiveness before application, thus optimising patient care. The increasing adoption of DTs technology in healthcare is evident across both public and private sectors. The Ecosystem for DTs in Healthcare (EDITH), a Coordination and Support Action (CSA) involving 19 partners, including key contributors from the United Kingdom, is exploring the integration of virtual human twins in a multidisciplinary approach. The private sector is also embracing this innovation, with Philips USA developing a Heart Model application that provides cardiologists with interactive 3D models for surgical planning. Similarly, companies like Siemens Healthineers and GE Healthcare, as highlighted by Copley Carol, are advancing the development of similar products to enhance personalised healthcare [4,5].

While the concept of DTs has been well established in industrial design, its application in healthcare introduces new ethical challenges. Unlike non-living objects, the human body presents complexities that are not easily captured by engineering models [18]. Our current understanding of genetics and disease mechanisms remains incomplete, making the accurate prediction of health outcomes through DTs challenging. This complexity is compounded by the fact that the engineering mindset, which often prioritises efficiency and functionality, may overlook critical human factors such as personal values, cultural contexts, and social environments.

To date, there has been a lack of comprehensive reviews on the ethical risks associated with the use of DTs in personalised healthcare. Although concerns related to data privacy represent only one facet of the broader ethical landscape [19]. The algorithmic biases embedded within predictive models could exacerbate existing inequalities, potentially leading to biased healthcare decisions. DTs might deepen social disparities, particularly for disadvantaged populations who may lack access to the necessary data collection devices and technologies [18]. As the healthcare sector continues to integrate Digital Twin technologies, it is crucial to address these ethical concerns to ensure that the benefits of PDTs are equitably distributed and do not reinforce existing inequalities. These considerations have been central to the development of our models, ensuring that ethical principles guide our research and application of PDTs.

In a global context, PDTs hold significant promise for advancing personalised medicine. However, their development is fraught with challenges related to design, ethical considerations, and data management. The concerns highlight the need for a comprehensive approach that addresses these hurdles while maximising the potential benefits of PDTs [20-24].

This study delves deeper into the broader implications of personalised healthcare, drawing on these observations that may be helpful in proposing strategies for the responsible and effective implementation of PDTs.

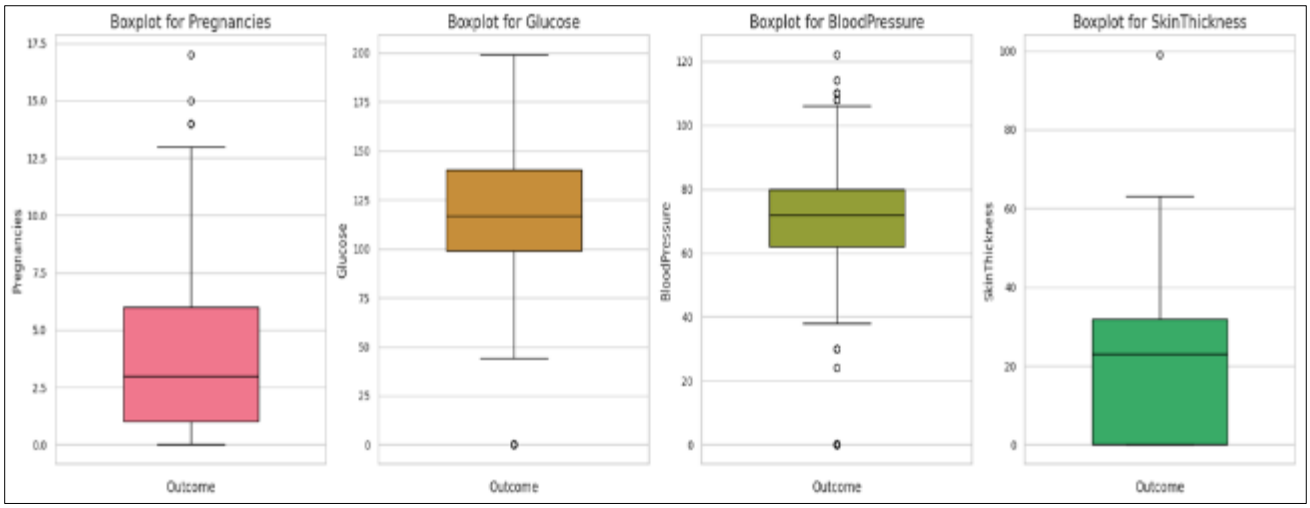
1.1. Developing Key Machine Learning Models

In the field of Machine Learning (ML), it is well-documented that identifying an optimal algorithm for a specific task is highly dependent on several factors, including the characteristics of the underlying dataset and the choice of performance metrics. Furthermore, the effectiveness of any chosen algorithm is often influenced by the tuning of hyperparameters, which are external to the model and control the training process but cannot be directly inferred from the data [25,26]. Additionally, there exists a prevalent misconception that more complex models inherently perform better on all datasets. While this may hold true for tasks involving unstructured data, such as biomedical imaging and video analysis, where Deep Learning (DL) models are advantageous due to their ability to automatically learn and hierarchically combine key spatial and temporal features [27,28] it is not necessarily the case for structured data. In such instances, interpretable models like logistic regression can achieve performance comparable to more sophisticated methods, such as extreme gradient boosting or neural networks [29,30]. Moreover, complex models are often prone to overfitting, where the model learns noise that fails to generalise to new datasets, emphasising the need for robust cross-validation techniques and a clear separation between training and testing datasets to ensure the external validity of the models [31]. The Dataset 1, tested under consideration comprises medical records from 768 female patients of PIMA Indian heritage, all aged 21 years or older. It includes eight predictor variables: the number of pregnancies, glucose concentration, blood pressure, skin thickness, insulin level, body mass index (BMI), diabetes pedigree function, and age. The outcome variable in this dataset indicates the presence or absence of diabetes. This dataset is open-source and is frequently utilised for the development and evaluation of ML models aimed at predicting diabetes.

A comprehensive pre-processing and analysis pipeline was implemented to ensure data quality and optimise model performance. Initially, outliers were detected and removed using the Interquartile Range (IQR) method. Specifically, for each predictor variable (excluding the 'Outcome' variable), the IQR was calculated, and data points lying beyond 1.5 times the IQR below the first quartile or above the third quartile were excluded from the dataset. This step was taken to eliminate anomalous observations that could potentially distort the analysis.

Following the outlier removal, missing values in variables such as 'SkinThickness', 'Insulin', 'Age', and 'BMI' were addressed by substituting zero values with the median of the corresponding columns. This imputation method was employed to mitigate the influence of missing or inaccurately recorded data, thereby resulting in a more robust dataset for analysis.

Exploratory Data Analysis (EDA) was subsequently conducted, which involved the generation of box plots, Figure 1.



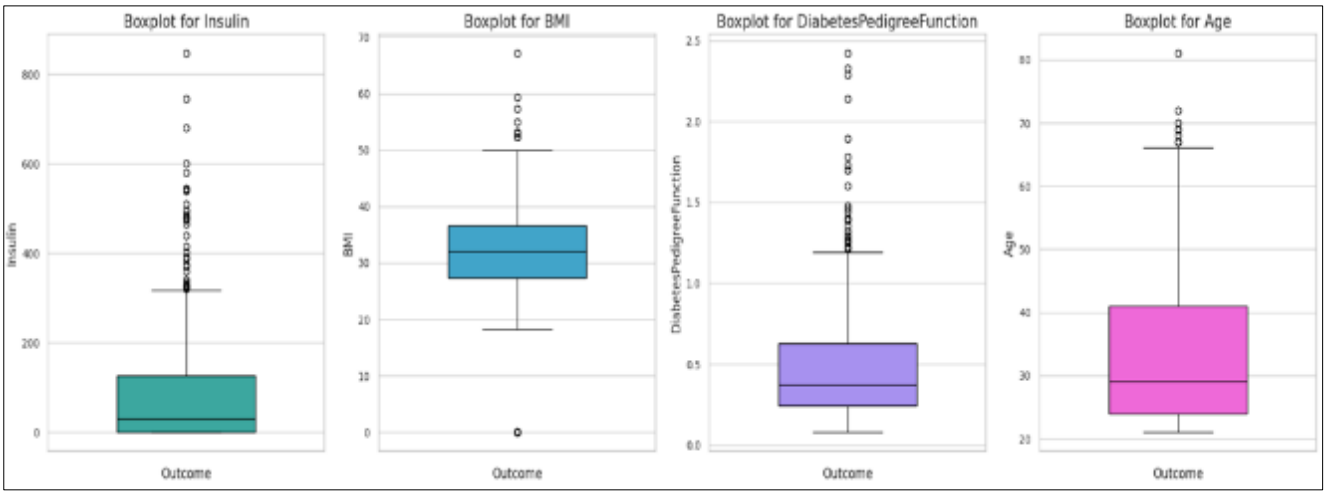


Figure 1 Visualising Outliers using Boxplot for Insulin

Histograms, Figure 2 (a, b), were generated for each predictor variable to visualise the data distributions and identify any remaining anomalies.

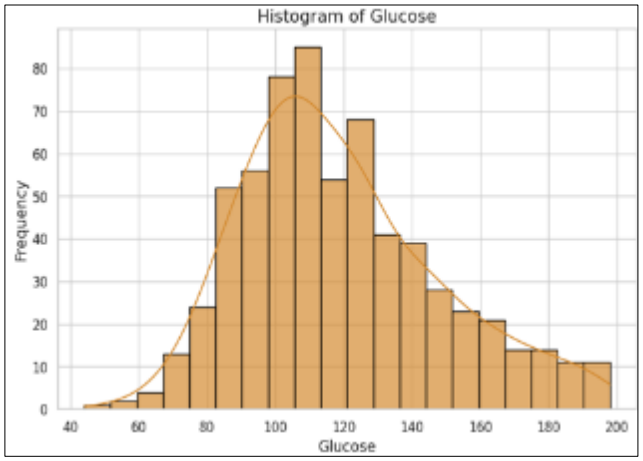


Figure 2 a) Histogram for Glucose

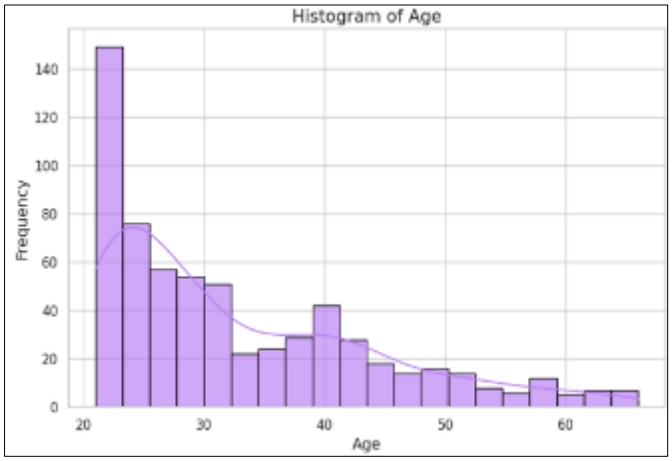


Figure 2 b) Histogram for Age

Additionally, a pair plot, Figure 3 was created to investigate the relationships between the predictor variables and their distributions across the 'Outcome' variable, providing insights into potential correlations and interactions that could inform the predictive modeling process.

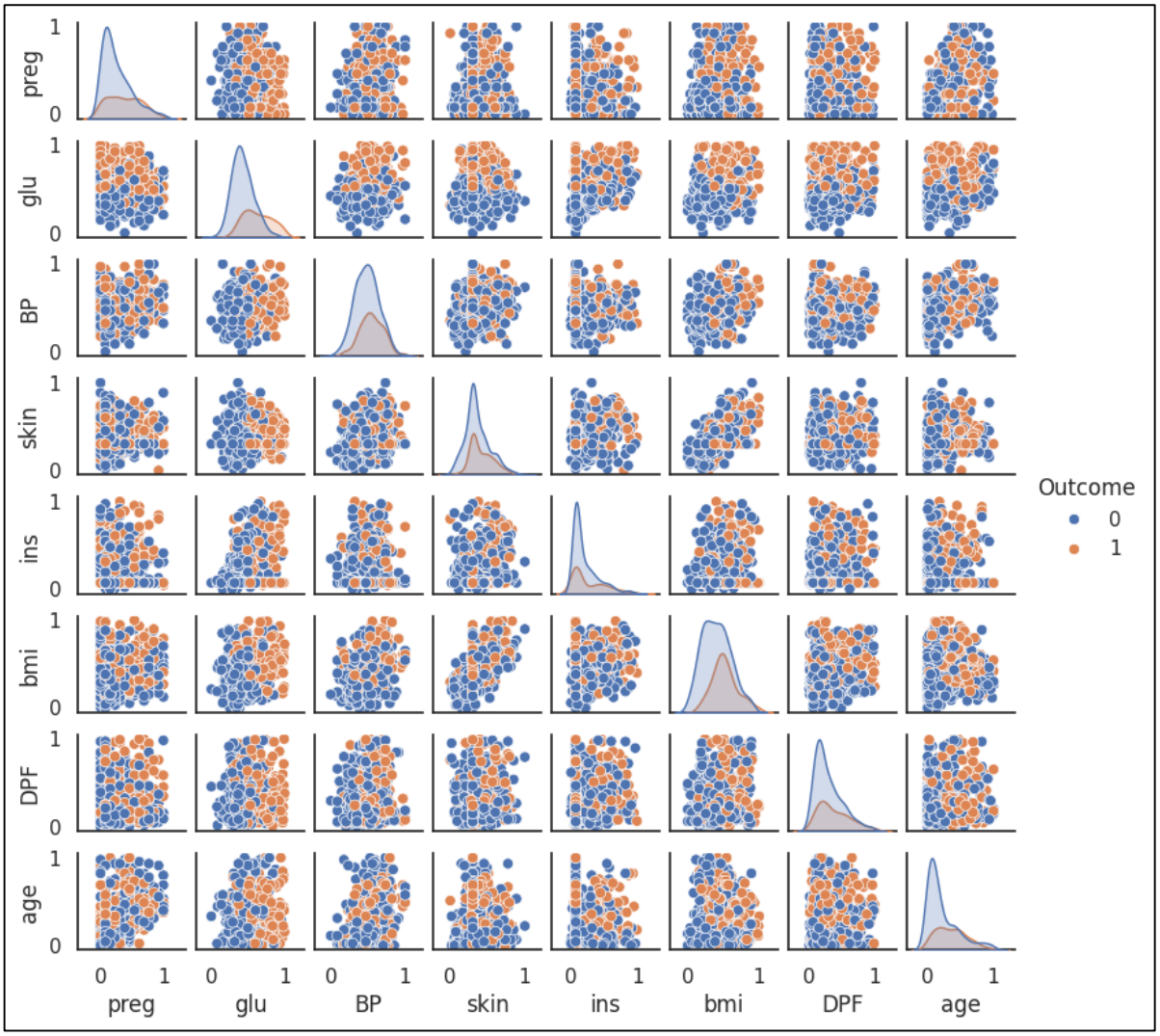


Figure 3 Pair plots

Figure 4 presents a heatmap utilised in our research to visualise the correlations between variables, thereby providing a comprehensive understanding of their interrelationships. By employing color gradients to indicate the strength and direction of these relationships, the heatmap facilitates the identification of key factors that influence diabetes outcomes. This visualisation is particularly valuable for feature selection, as it highlights variables that are strongly correlated with the outcome and with each other, thus helping to mitigate potential multicollinearity issues. Additionally, the heatmap offers a clear and intuitive overview of the dataset, which supports informed decision-making in model development and enhances the interpretability of the results.

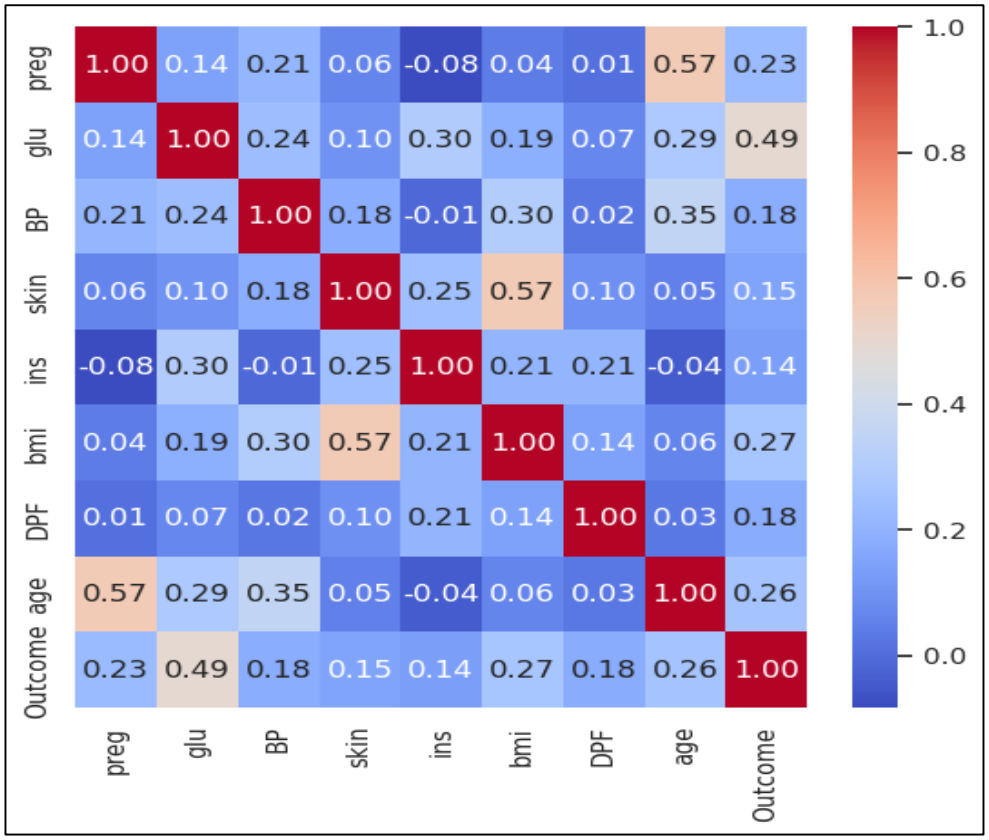


Figure 4 Heatmap of Correlation Matrix

Subsequently, several ML models were trained and evaluated using this dataset. The first model applied was Logistic Regression, with the dataset divided into training and testing sets using an 80-20 split. The model's performance was assessed using metrics such as accuracy score, confusion matrix, and classification report.

Following the Logistic Regression, a Gradient Boosting Correlation (GBC) was trained and evaluated, offering a comparative analysis of model performance on the standardised dataset. The GBC demonstrated superior performance compared to the Logistic Regression model. Additionally, a neural network was implemented using TensorFlow, consisting of an input layer with 8 nodes, two hidden layers with 64 and 32 nodes respectively, and a single output node. The model was trained using binary cross-entropy loss over 10 epochs, with accuracy and loss monitored across epochs for both the training and validation datasets. Furthermore, hyperparameter optimisation was conducted on a LightGBM model using RandomisedSearchCV. The best parameters were selected based on the highest accuracy score, and the model's performance was subsequently evaluated using a confusion matrix, classification report, and accuracy score, providing a comprehensive assessment of its predictive capabilities.

The Dataset 2 tested is the UCI Diabetes Open-source Dataset, collected through patient questionnaires at Sylhet Diabetes Hospital in Bangladesh and subsequently verified by medical professionals, is designed to predict early-stage diabetes. This dataset includes 17 attributes, such as age, sex, and various symptoms like polyuria, polydipsia, and sudden weight loss, alongside a target variable indicating diabetes status (positive/negative).

Histograms were generated for all features to visualise their distributions, and a correlation heatmap, Figure 5 was produced to explore the relationships among the features. A critical component of the preprocessing for the second dataset was feature selection. Features such as 'delayed healing', 'Itching', and 'Obesity' were removed based on their correlation with other variables and their impact on model performance, as determined through correlation analysis and initial model training.

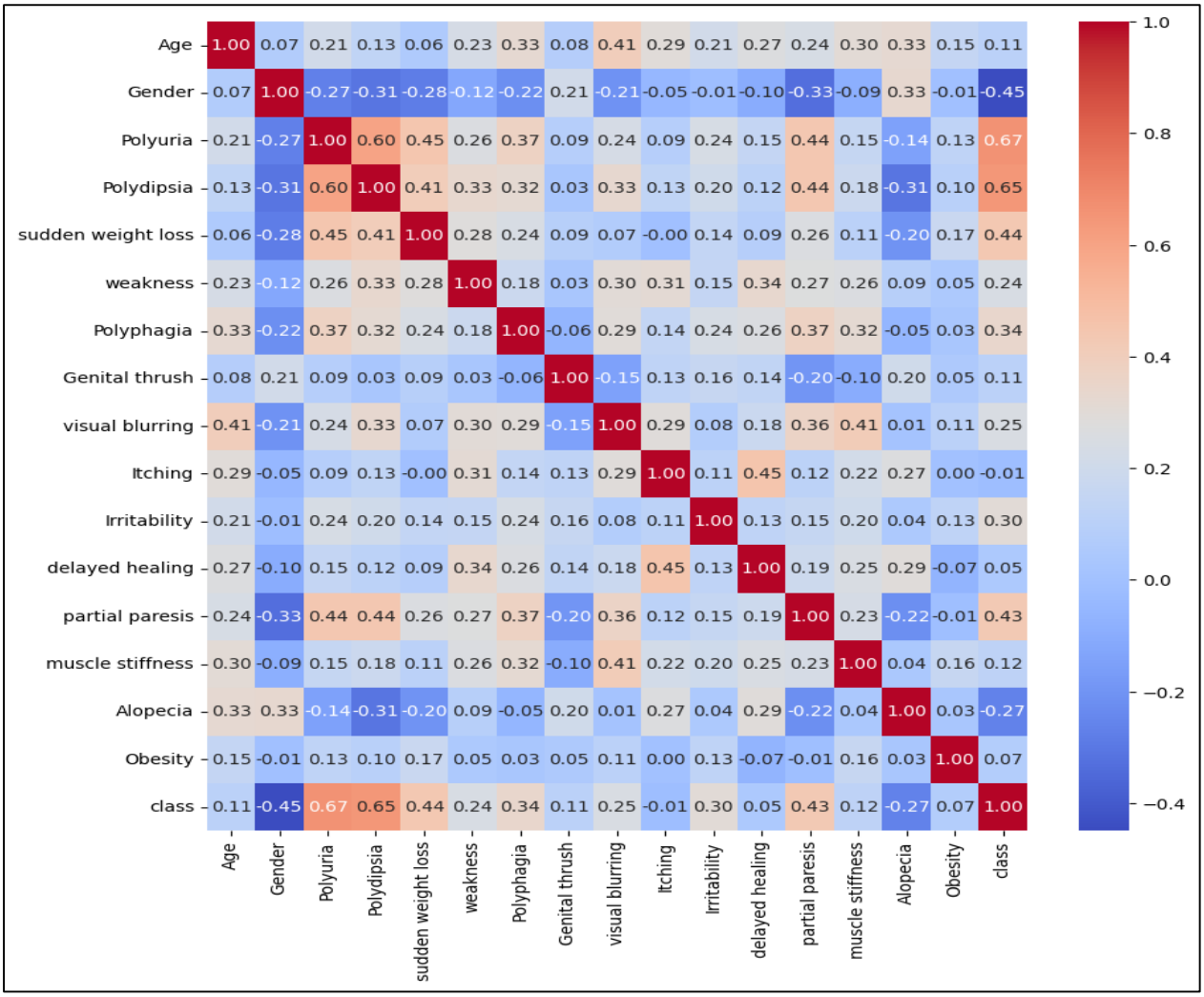


Figure 5 Correlation Heatmap

Multiple models were trained and evaluated on this dataset. The initial model, Logistic Regression, was assessed using accuracy, confusion matrix, and classification report metrics. A Random Forest Classifier, trained with a maximum depth of 2, was also evaluated using the same metrics. The GBC was subsequently trained and demonstrated superior performance compared to the other models. Furthermore, the impact of feature selection was analysed by training models on both the complete feature set and a reduced feature set (excluding 'delayed healing', 'Itching', and 'Obesity'). Performance metrics from these models were compared to assess the influence of feature selection on overall model accuracy.

2. Discussion

The results section compares the performance of various ML models on two distinct open-source diabetes datasets: the PIMA Indian Diabetes Dataset and the UCI Diabetes Dataset. The evaluation metrics used to compare model performance include accuracy, confusion matrix, and classification report, which provide insights into the models' predictive capabilities and robustness. This section highlights the key findings and draws comparisons between the datasets to understand the impact of different preprocessing techniques and feature selection on model performance.

For Dataset 1, the PIMA Indian Diabetes Dataset, the Logistic Regression model achieved an accuracy of 78%. The confusion matrix showed a balanced performance with a moderate number of false positives and false negatives, and the classification report indicated consistent precision, recall, and F1-scores across both classes, reflecting reasonable model performance, Figure 6. Logistic Regression is a widely used statistical method for binary classification, frequently applied in diabetes management and prediction [32]. It models the probability that a given input belongs to a particular class by utilising the logistic function, which maps input values to probabilities ranging between 0 and 1. The model

optimises its parameters to identify the best-fit line that separates the classes. Although it is a relatively simple method, Logistic Regression is powerful and interpretable, making it particularly useful for tasks such as disease diagnosis, including diabetes [33]. Logistic Regression can be extended to handle multiclass classification problems and incorporates regularisation techniques to prevent overfitting and manage multicollinearity, ensuring more robust model performance in clinical settings [34,35].

The GBC outperformed Logistic Regression with an accuracy of 82%, demonstrating fewer false positives and false negatives, and yielding higher precision, recall, and F1-scores, particularly for the positive class, Figure 6. The GBC is a sophisticated ensemble learning technique that constructs predictive models sequentially. Each successive tree in the ensemble aims to correct the errors made by the preceding trees by minimising a specified loss function via gradient descent. This iterative approach enables the GBC to effectively handle various types of data, including numerical, categorical, and incomplete datasets, making it highly suitable for both classification and regression tasks [36,37]. The GBC has demonstrated substantial utility in improving predictive accuracy for diabetes-related outcomes. For instance, its application in diabetes risk prediction models has shown significant advancements in identifying at-risk individuals through accurate risk stratification [38]. Moreover, the technique's adaptability to diverse data types enhances its robustness and generalizability across different datasets [39]. However, the effectiveness of GBC is contingent upon meticulous hyperparameter tuning and regularisation to mitigate overfitting, a common challenge in complex predictive modelling [40]. Thus, while the GBC offers powerful predictive capabilities, its optimal performance in diabetes research necessitates careful consideration of model parameters and validation procedures to ensure reliable and generalisable outcomes [41].

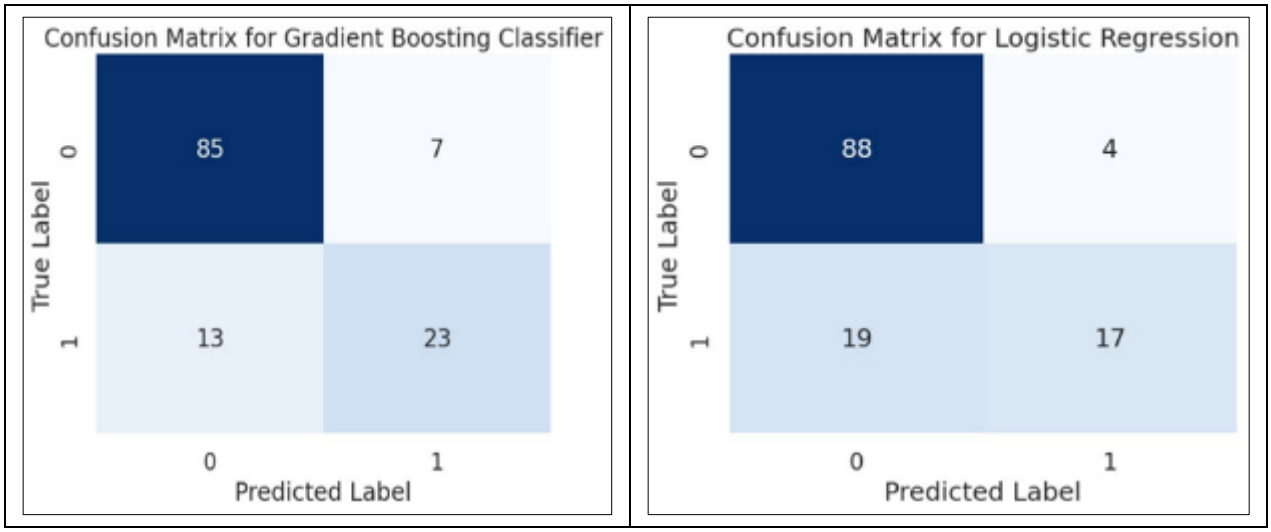


Figure 6 Confusion matrix for GBC and Logistic Regression for PIMA Dataset

The Neural Network also performed well with an accuracy of 80%, showing a comparable confusion matrix to the GBC but with slightly more false positives. Neural networks are a class of ML models designed to emulate the architecture and functioning of the human brain. These models consist of interconnected layers of nodes, or neurons, and are particularly adept at processing large-scale datasets and capturing complex patterns [42]. Neural network architectures typically include an input layer, one or more hidden layers, and an output layer. During the training process, connections between neurons, or weights, are adjusted via backpropagation to minimise prediction errors [43]. Neural Networks have proven highly effective in managing and analysing complex data, such as patient records, genetic information, and imaging data [44]. The introduction of activation functions within these networks allows for the modelling of non-linear relationships, which is crucial for identifying intricate patterns in diabetes-related data [45]. Advanced neural network architectures, including Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), are particularly valuable in this domain. CNNs are employed for analysing high-dimensional data such as medical imaging, where they have been used to enhance diagnostic accuracy in diabetic retinopathy [46]. RNNs, on the other hand, are useful for sequential data, such as time-series health records, enabling improved prediction of diabetes progression [47].

In the case of Dataset 2, the UCI Diabetes Dataset, the Logistic Regression model achieved an accuracy of 91%, which was higher than Dataset 1. The confusion matrix revealed a higher number of false positives and false negatives, indicating the dataset's complexity. The classification report showed lower precision and recall compared to Dataset 1, suggesting that the UCI dataset might be less challenging for Logistic Regression. The Random Forest Classifier

improved performance with an accuracy of 90%, showing fewer false positives and false negatives compared to Logistic Regression. The classification report for the Random Forest Classifier indicated better precision, recall, and F1-scores, demonstrating the advantage of ensemble methods. The Random Forest Classifier is an ensemble learning method widely used for both classification and regression tasks, including diabetes prediction. This technique involves the construction of multiple decision trees during the training process. Each tree is trained on a randomly chosen subset of the data and features, which enhances model diversity and reduces the likelihood of overfitting [48]. The Random Forest Classifier has been shown to be highly effective due to its ability to handle a large number of input variables and provide valuable insights into feature importance [49]. This method's robustness in managing noisy and incomplete data makes it particularly suitable for healthcare applications, where it has been used to predict diabetes risk and outcomes with high accuracy and interpretability [50]. The Random Forest approach not only improves predictive performance but also offers a clear understanding of the contribution of each feature to the model, which is critical for clinical decision-making and personalised medicine [49]. Its efficacy in processing complex datasets and its ability to generate reliable predictions underscore its significant role in modern healthcare analytics.

The Gradient Boosting Classifier, with an accuracy of 96%, showed similar performance improvements, with a confusion matrix indicating fewer false positives and false negatives and a classification report reflecting high precision and recall, particularly for the positive class, Figure 7.

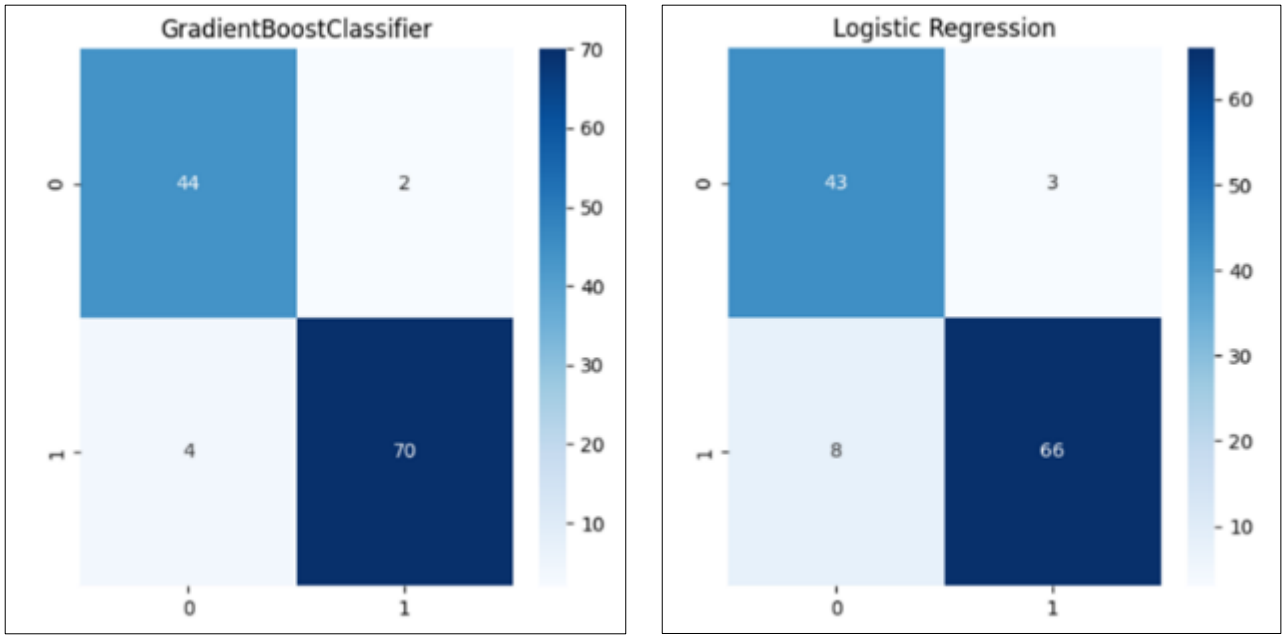


Figure 7 Confusion matrix for GBC and Logistic Regression for UCI Dataset

Comparative analysis reveals several insights. The GBC consistently achieved the highest accuracy across both datasets, with 82% for Dataset 1 and 80% for Dataset 2. Logistic Regression performed slightly better on Dataset 1 with an accuracy of 78%, compared to 75% on Dataset 2, reflecting the impact of extensive preprocessing in the PIMA dataset. Both datasets showed that ensemble methods like Gradient Boosting and Random Forest reduced the number of false positives and false negatives compared to Logistic Regression, a trend more pronounced in the UCI dataset, indicating its higher complexity and the need for robust models. Precision, recall, and F1-scores were generally higher for ensemble methods across both datasets. The PIMA dataset benefited significantly from thorough preprocessing, leading to better performance metrics for all models. In contrast, the UCI dataset, despite its inherent complexity, still showed notable improvements with Gradient Boosting and Random Forest models. The extensive preprocessing applied to the PIMA dataset, including outlier removal, handling missing values, and feature scaling, resulted in more stable and accurate model performance. The UCI dataset, which focused on encoding and feature selection, also demonstrated improved model performance, particularly with advanced models like 'Gradient Boosting.'

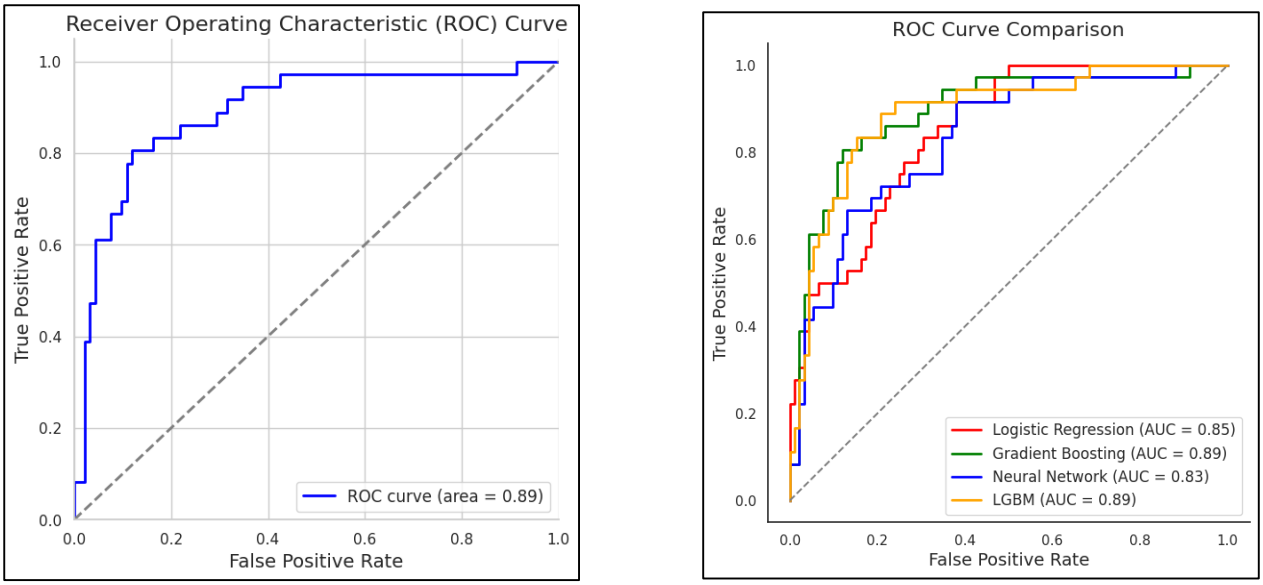


Figure 8 Receiver Operating Characteristic curve (ROC) for True-Positive and False positive rate at various threshold

The Receiver Operating Characteristic (ROC) curve, depicted in Figure 8, was pivotal in assessing the performance of machine learning models within the context of personalised healthcare. By plotting the true positive rate against the false positive rate across various thresholds, the ROC curve visually illustrates the trade-offs between sensitivity and specificity. This approach allowed for a comparative analysis of models such as Logistic Regression, Random Forest, Gradient Boosting, and Neural Networks in accurately distinguishing patient conditions. The area under the curve (AUC) served as a succinct performance metric, guiding the optimisation of algorithms to improve predictive accuracy in clinical settings. Consequently, this analysis supports the development of more tailored and effective healthcare interventions.

The application of decision trees in healthcare demonstrates considerable promise, particularly for enhancing predictive analytics and personalised treatment strategies. However, several ethical considerations must be addressed to ensure responsible data use. These include the importance of clear data visualisation, accessibility, the ability to easily add or remove data sources to address privacy concerns or correct errors, and the integration of these systems into clinical workflows [51]. A modular architecture for DTs is recommended to address these challenges. In this framework, each biological process represented within the digital twin is encapsulated as a separate module, interacting through a central data structure. This design minimises direct dependencies between modules and facilitates the easy addition or removal of data sources, enhancing both privacy protection and system extensibility [52].

Furthermore, the establishment of digital twin consortia, such as the Swedish Digital Twin Consortium (SDTC), is vital for developing standardized methods and interoperability protocols for DTs. Such collaborations among industry, government, academia, and practitioners are essential for advancing the field and ensuring consistent application across different contexts [53]. To further support the development and application of DTs in healthcare, a dedicated, multi-purpose human/medical digital twin software development kit (SDK) is proposed. This specialized SDK would complement existing generic tools by optimising them specifically for healthcare applications, thereby enhancing the utility and effectiveness of DTs in clinical settings [54].

3. Ethical Concerns

The ethical considerations surrounding PDTs in healthcare are complex and multifaceted, necessitating careful examination at multiple levels. Key ethical issues include data control and privacy, where concerns revolve around data ownership, preventing misuse, and safeguarding individual privacy within PDTs. Algorithmic bias is another critical concern, as PDTs have the potential to perpetuate existing biases, thereby impacting healthcare decisions. The degree of technical autonomy and control is also a significant issue, raising questions about individual control over PDTs and liability for the actions of these systems [55]. End-of-life considerations further complicate the ethical landscape, with questions about the fate of a PDT after an individual's death [56]. DTs using nanoparticle-based measurements for body functions could revolutionise health services, they also pose significant risks such as body hacking, identity theft, new forms of cybercrime, and even the potential militarisation of basic resources [57], while the importance of preserving

the child's role in pediatric care when using DT systems [58-60]. Furthermore, the interaction between the represented persons and their simulations raises questions regarding privacy, autonomy, and the control individuals have over their DTs [61, 62]. These ethical challenges need to be addressed to ensure responsible and ethical use of PDTs in healthcare. The integration of Large Language Models (LLMs) with PDTs in healthcare offers significant potential for personalised medicine and improved patient care, as it enables more accurate and tailored digital representations of individuals, leading to more effective interventions. However, this combination also raises important challenges, including privacy concerns, the need for transparency in data use, and potential negative impacts on doctor-patient communication, which must be carefully addressed to ensure the responsible and equitable use of these technologies [63-66].

4. Conclusion

In conclusion, the findings emphasize the critical role of pre-processing and feature engineering in optimising model performance. Analysis of the PIMA Indian Diabetes Dataset, which underwent comprehensive preprocessing, generally yielded higher accuracy and improved performance metrics. Conversely, the UCI Diabetes Dataset, which presents greater challenges, still demonstrated substantial benefits from sophisticated modelling techniques, particularly ensemble methods. Among the models evaluated, the GBC emerged as the most robust, exhibiting superior performance across both datasets. This underscores its effectiveness in managing complex and heterogeneous data. Future research should investigate additional preprocessing techniques and further explore ensemble modeling strategies to enhance predictive accuracy and model robustness.

PDTs have the potential to revolutionise personalised healthcare by using G-AI technologies to create virtual models of patients based on their health data. These DTs can be used for various purposes such as drug testing, predictive analysis, disease modelling, and lifestyle improvement. However, it is important to ensure that the use of G-AI in healthcare is guided by human expertise and judgement, and that the accuracy, reliability, and ethical considerations of these technologies are continually assessed. More of empirical literature and equitable principles should be highlighted to further appropriate the frameworks of ethics and norms in DTs functions and usage. In conclusion, human digital twins represent a powerful tool for personalised medicine, but careful consideration must be given to design, ethical, and data management challenges to ensure its responsible and equitable implementation.

Compliance with ethical standard

Acknowledgments

Authors acknowledge the contribution of all the authors.

Disclosure of conflict of interest

No conflict of interest to be disclosed.

Open-Source Datasets

- PIMA Indian Diabetes Dataset
- UCI Diabetes Dataset

References

[1] Habibzadeh H, Dinesh K, Shishvan OR, Boggio-Dandry A, Sharma G, Soyata T. A survey of healthcare internet of things (HIoT): A clinical perspective. IEEE Access. 2020; 8:215-234.

[2] Yang Y, Wang H, Jiang R, Wang X, Li Y, Liu X. A review of IoT-enabled mobile healthcare: Technologies, challenges, and future trends. IEEE Internet Things J. 2022;7(1):53-71.

[3] Gawel DR, Serra-Musach J, Lilja S, Aagesen J, Arenas A, Asking B, Bengnér M, Björkander J, Biggs S, Ernerudh J, Hjortswang H. A validated single-cell-based strategy to identify diagnostic and therapeutic targets in complex diseases. Genome medicine. 2019 Dec;11:1-25.

[4] Copley C. Medtech firms get personal with digital twins. Reuters. 2018 Aug 23. Available from: https://www.reuters.com/article/us-healthcare-medical-technology-ai-insi-idUSKCN1LG0S0

[5] Philips-USA. Dynamic HeartModel A.I. (Anatomical Intelligence) for additional insight. Philips-USA. 2018. Available: https://www.usa.philips.com/healthcare/resources/feature-detail/ultrasound-heartmodel

[6] Rubeis G. Hyperreal Patients. Digital Twins as Simulacra and their impact on clinical heuristics. InMedizin–Technik–Ethik: Spannungsfelder zwischen Theorie und Praxis 2023 Jan 1 (pp. 193-207). Berlin, Heidelberg: Springer Berlin Heidelberg.

[7] Popa EO, van Hilten M, Oosterkamp E, Boga MJ, van Beek FE, Vossen HR, Spreeuwenberg P, Essink-Bot ML, Prainsack B, Essink-Bot ML, Helbing D. The use of digital twins in healthcare: Socio-ethical benefits and socio-ethical risks. Life Sci Soc Policy. 2021;17(1):1-15.

[8] Bruynseels K, Santoni de Sio F, van den Hoven J. Digital Twins in healthcare: Ethical implications of an emerging engineering paradigm. Front Genet. 2018; 9:31.

[9] Indiana University. Digital twins could lead to more proactive personalized medicine, researchers say. Medical Xpress. 2021 Mar 3. Available from: https://medicalxpress.com/news/2021-03-digital-twins-proactive-personalized-medicine.html

[10] Schwartz SM, Wildenhaus K, Bucher A, Byrd B. Digital twins and the emerging science of self: Implications for digital health experience design and "small" data. Front Comput Sci. 2020; 4:31.

[11] Xames MD, Topcu TG. A Systematic Literature Review of Digital Twin Research for Healthcare Systems: Research Trends, Gaps, and Realization Challenges. IEEE Access. 2024 Jan 3.

[12] Qian S, Ugurlu D, Fairweather E, Strocchi M, Toso LD, Deng Y, Plank G, Vigmond E, Razavi R, Young A, Lamata P. Developing Cardiac Digital Twins at Scale: Insights from Personalised Myocardial Conduction Velocity. medRxiv. 2023 Dec 5:2023-12.

[13] Masison J, Beezley J, Mei Y, Ribeiro H, Knapp AC, Vieira LS, Watanabe M, Pasquale MD, Wheeler DW. A modular computational framework for medical digital twins. PubMed. 2021;339(3):1865-1877.

[14] Rivera LF, Jiménez M, Angara P, Villegas NM, Tamura G, Müller HA. Towards continuous monitoring in personalized healthcare through digital twins. InProceedings of the 29th annual international conference on computer science and software engineering 2019 Nov 4 (pp. 329-335).

[15] Balasubhramanyam A, Ramesh R, Sudheer R, Honnavalli PB. Revolutionizing Healthcare: A Review Unveiling the Transformative Power of Digital Twins. IEEE Access. 2024 May 13.

[16] Márquez FP, Korhan O. Digital Twin Technology-Fundamentals and Applications: Fundamentals and Applications. BoD–Books on Demand; 2023 Dec 20.

[17] Bruynseels K., Santoni de S. F., van den H. J. Ethical implications of digital twins in healthcare: A critical perspective. Front Genet. 2018; 9:31.

[18] Kamel Boulos MN, Zhang P. Digital twins: from personalised medicine to precision public health. Journal of personalized medicine. 2021 Jul 29;11(8):745.

[19] Nasir S, Khan RA, Bai S. Ethical framework for harnessing the power of ai in healthcare and beyond. IEEE Access. 2024 Feb 26;12:31014-35.

[20] Huang PH, Kim KH, Schermer M. Ethical issues of digital twins for personalized health care service: preliminary mapping study. Journal of Medical Internet Research. 2022 Jan 31;24(1):e33081.

[21] Xames MD, Topcu TG. A Systematic Literature Review of Digital Twin Research for Healthcare Systems: Research Trends, Gaps, and Realization Challenges. IEEE Access. 2024 Jan 3.

[22] Kerckhove D. The personal digital twin: Ethical and philosophical considerations. Philos Trans R Soc A. 2021;379(2197):20200367.

[23] Helbing D., Sánchez-Vaquerizo J. Digital Twins: Potential, ethical issues, and limitations. arXiv. 2022. Available from: https://arxiv.org/ftp/arxiv/papers/2208/2208.04289.pdf

[24] Ahmadi-Assalemi G, Al-Khateeb H, Maple C, Epiphaniou G, Alhaboby ZA, Alkaabi S, Tchorek-Dombrowska J. Digital twins for precision healthcare. In: Maple C, Epiphaniou G, editors. Cyber-Physical Systems for Next-Generation Healthcare. Cham: Springer; 2020. p. 123-136.

[25] Snoek, J., Larochelle, H., & Adams, R. P. (2012). Practical Bayesian optimization of machine learning algorithms. In Advances in Neural Information Processing Systems (pp. 2951-2959).

[26] Bergstra, J., Bardenet, R., Bengio, Y., & Kégl, B. (2011). Algorithms for hyper-parameter optimization. In Advances in Neural Information Processing Systems (pp. 2546-2554).

[27] LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436-444.

[28] Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. Nature, 542(7639), 115-118.

[29] Chen T, Guestrin C. Xgboost: A scalable tree boosting system. InProceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining 2016 Aug 13 (pp. 785-794).

[30] Weng SF, Reps J, Kai J, Garibaldi JM, Qureshi N. Can machine-learning improve cardiovascular risk prediction using routine clinical data?. PloS one. 2017 Apr 4;12(4):e0174944..

[31] Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. MIT Press.

[32] Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). Applied Logistic Regression (3rd ed.). Wiley.

[33] Pokala VS, Kumar NS. Analysis and Comparison for Prediction of Diabetic among Pregnant Women using Innovative K-Nearest Neighbor Algorithm over Logistic Regression with Improved Accuracy. Cardiometry. 2022 Dec 1(25):949-55.

[34] Cox, D. R. (1958). The regression analysis of binary sequences. Journal of the Royal Statistical Society: Series B (Methodological), 20(2), 215-242.

[35] Kianifard F. Logistic regression: A self-learning text.

[36] Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. The Annals of Statistics, 29(5), 1189-1232.

[37] Chen, J., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785-794.

[38] Pencina, M. J., D'Agostino, R. B., & Vasan, R. S. (2008). Evaluating the added predictive ability of a new marker: From area under the ROC curve to reclassification and beyond. Statistics in Medicine, 27(2), 157-172.

[39] Tulchhia A, Rathore M. A Survey On Overall Survival Prediction Of Brain Tumor Patients Using Machine Learning And Gradient Boosting Techniques. Webology (ISSN: 1735-188X). 2021;18(4).

[40] Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction (2nd ed.). Springer.

[41] Caruana R, Karampatziakis N, Yessenalina A. An empirical evaluation of supervised learning in high dimensions. InProceedings of the 25th international conference on Machine learning 2008 Jul 5 (pp. 96-103).

[42] LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436-444.

[43] Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. Nature, 323(6088), 533-536.

[44] Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., & Blau, H. M. (2019). Dermatologist-level classification of skin cancer with deep neural networks. Nature, 542(7639), 115-118.

[45] Hinton, G. E., Osindero, S., & Teh, Y. W. (2012). A fast learning algorithm for deep belief nets. Neural Computation, 18(7), 1527-1554

[46] Gulshan, V., Peng, L., Coram, M., Stumpe, M. C., & Wu, D. (2016). Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA, 316(22), 2402-2410.

[47] Cho K. On the Properties of Neural Machine Translation: Encoder-decoder Approaches. arXiv preprint arXiv:1409.1259. 2014.

[48] Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5-32.

[49] Liaw A, Wiener M. Classification and regression by randomForest. R news. 2002 Dec 3;2(3):18-22.

[50] Haq AU, Li J, Memon MH, Nazir S, Ahmad S, Ali A. A new intelligent approach for effective recognition of diabetes in the IoT e-healthcare environment.

[51] Himmler, J., Hübner, M., & Marzbani, H. (2020). Ethical considerations in the use of digital twins for health care. Journal of Biomedical Informatics, 104, 103388.

[52] Pereira, G., Figueiredo, J., & Ribeiro, A. (2021). Modular architectures for digital twins: Design principles and challenges. Advanced Intelligent Systems, 3(5), 2100012.

[53] Kang, J., Li, Y., & Tan, X. (2022). Standardization and interoperability of digital twin technologies: A review. IEEE Access, 10, 9876-9887.

[54] Zhang, Z., Liu, J., & Liu, H. (2023). Development and application of a digital twin software development kit for healthcare. Journal of Healthcare Engineering, 2023, 154782.

[55] Braun M. Ethics of digital twins: four challenges. J Med Ethics. 2021;48(9):579-580.

[56] Huang PH, Kim KH, Schermer M. Ethical Issues of Digital Twins for Personalized Health Care Service: Preliminary Mapping Study. J Med Internet Res. 2022;24(1)

[57] Helbing D, Argota Sanchez-Vaquerizo J. Digital twins: Potentials, limitations, and ethical issues. Preprint, version 0.17. https://www. researchgate. net/publi catio. 2022(35857).

[58] Maeckelberghe E, Zdunek K, Marceglia S, Farsides B, Rigby M. The ethical challenges of personalized digital health. Frontiers in Medicine. 2023 Jun 19;10:1123863..

[59] Drummond D, Coulet A. Technical, ethical, legal, and societal challenges with digital twin systems for the management of chronic diseases in children and young people. Journal of Medical Internet Research. 2022 Oct 31;24(10):e39698.

[60] Tigard DW. Digital twins running amok? Open questions for the ethics of an emerging medical technology. Journal of medical ethics. 2021 Jun 1;47(6):407-8.

[61] Iqbal JD, Krauthammer M, Biller-Andorno N. The use and ethics of digital twins in medicine. Journal of Law, Medicine & Ethics. 2022 Sep;50(3):583-96.

[62] Mittelstadt B. Near-term ethical challenges of digital twins. Journal of medical ethics. 2021 Jun 1;47(6):405-6..

[63] Sahal, R.; Alsamhi, S.H.; Brown, K.N. Personal Digital Twin: A Close Look into the Present and a Step towards the Future of Personalised Healthcare Industry. *Sensors* **2022**, *22*, 5918.

[64] Wang Y, Fu T, Xu Y, Ma Z, Xu H, Du B, Lu Y, Gao H, Wu J, Chen J. Twin-gpt: Digital twins for clinical trials via large language model. ACM Transactions on Multimedia Computing, Communications and Applications. 2024 Apr 1..

[65] Hong Y, Wu J, Morello R. Llm-twin: Mini-giant model-driven beyond 5g digital twin networking framework with semantic secure communication and computation. Scientific Reports. 2024 Aug 17;14(1):19065.

[66] Zalake M. Doctors' perceptions of using their digital twins in patient care. Scientific Reports. 2023 Dec 7;13(1):21693.