ISSN: 2454-9940

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT

E-Mail :
editor.ijasem@gmail.com
editor@ijasem.org

www.ijasem.org

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT

ISSN 2454-9940

www.ijasem.org

Vol 19, Issue 2, 2025

APPLICATION OF BIG DATA PROCESSING AND ARTIFICIAL INTELLIGENCE IN PRECISION MEDICINE

Uddaraju Laxmi Sravya

21N81A6725

Computer Science and Engineering (Data-Science)

Sphoorthy Engineering College,

Hyderabad,501510

uddarajulaxmisravya@gmail.com

Palla Mrudula

21N81A6709

Computer Science and Engineering (Data-Science)

Sphoorthy Engineering College, Nadergul,

Hyderabad,501510

pallamrudula49@gmail.com

Gurram Rohit Reddy

21N81A6743

Computer Science and Engineering (Data-Science)

Sphoorthy Engineering College,

Hyderabad,501510

reddyrohit975@gmail.com

Deegunti Deekshitha

22N85A6701

Computer Science and Engineering (Data-Science)

Sphoorthy Engineering College, Nadergul,

Hyderabad,501510

deekshithadeegunti@gmail.com

Dr. Karanam Ramesh Rao

Professor

Computer Science and Engineering (Data-Science)

Sphoorthy Engineering College,

Nadergul, Hyderabad,501510

rameshrao@sphoorthyengg.ac.in

ABSTRACT: Precision medicine, also known as personalized medicine, represents a transformative approach in healthcare by offering customized treatment strategies based on individual patient profiles. Unlike conventional methodologies, it leverages personal data such as genetic makeup, environment, and lifestyle factors to optimize treatment outcomes. With the increasing availability of big data and advancements in artificial intelligence (AI), it is now possible to process and analyze vast volumes of complex patient data. This project investigates how AI and big data technologies, particularly Hadoop and machine learning algorithms like Logistic Regression and XGBoost, contribute to improving diagnostic precision and personalized healthcare—especially in the context of neurological

diseases like Alzheimers. Despite the promise, challenges in data quality, privacy, and infrastructure must be addressed to realize the full potential of precision medicine. The system developed here emphasizes model transparency and clinical utility through tools like SHAP (SHapley Additive exPlanations), aiming to support decision-making in real-world medical settings.

1. INTRODUCTION

Alzheimer's Disease (AD) is a chronic neurodegenerative condition that affects millions worldwide, leading to a progressive decline in memory, cognitive function, and the ability to perform everyday tasks. It is currently incurable, and most interventions focus on symptom management rather than disease reversal. As the global population continues to age, the number of individuals affected by AD is projected to increase significantly, placing immense pressure on healthcare systems, caregivers, and society at large.

A major limitation in current diagnostic and treatment strategies for Alzheimer's is the reliance on a generalized approach. Conventional medicine typically applies the same treatment plans to broad patient groups, regardless of individual variability in genetic, clinical, or lifestyle factors. This "one-size-fits-all" model often results in delayed diagnosis, missed early intervention opportunities, and inconsistent treatment effectiveness across patient populations.

To overcome these limitations, the field of precision medicine also known as personalized medicine has emerged. Precision medicine aims to tailor medical care to the individual characteristics of each patient. By integrating data from various sources such as medical history, laboratory results, and other clinical features, healthcare professionals can make better-informed decisions that are customized to the unique profile of each patient.

The recent rise of artificial intelligence (AI) and machine learning (ML) has revolutionized precision medicine by enabling the analysis of large and complex datasets. AI techniques can extract meaningful patterns and insights from data that would otherwise be difficult to interpret manually. These insights support earlier detection, more accurate diagnosis, and personalized treatment recommendations, particularly in diseases like Alzheimer's where early intervention is crucial.

In this project, we focus on building a machine learning-based diagnostic system that predicts the presence of Alzheimer's Disease using structured clinical data. The system is designed using two primary classification models: Logistic Regression and XGBoost (Extreme Gradient Boosting). Logistic Regression is a widely used linear model known for its simplicity, interpretability, and effectiveness in binary classification tasks. XGBoost, on the other hand, is an advanced ensemble learning method that leverages gradient boosting techniques to deliver superior predictive performance, especially on structured data.

To ensure the model is not just accurate but also interpretable an essential requirement in healthcare we employ SHAP (SHapley Additive exPlanations). SHAP values quantify the impact of each input feature on the model's predictions, providing clinicians with a transparent understanding of why a particular diagnosis was suggested. This aspect of explainable AI

ISSN 2454-9940
www.ijasem.org
Vol 19, Issue 2, 2025

is critical for fostering trust and facilitating clinical decision-making.

The system is developed using Python and common data science libraries, including pandas, NumPy, scikit-learn, XGBoost, and SHAP. It includes data preprocessing, feature selection, model training and evaluation, and result visualization through tools such as confusion matrices, classification reports, and ROC curves.

By combining traditional and advanced machine learning techniques with a strong emphasis on interpretability, this work contributes to the advancement of precision medicine in the context of Alzheimer's Disease. It demonstrates how structured clinical data, when paired with robust AI tools, can support early and more personalized diagnosis ultimately leading to better patient outcomes and more informed healthcare strategies.

2. LITERATURE REVIEW

[1] Gupta, N. S., & Kumar, P. (2023). *Perspective of Artificial Intelligence in Healthcare Data Management: A Journey Towards Precision Medicine*. This study emphasizes the integration of artificial intelligence (AI) into healthcare data management to advance precision medicine. It highlights how AI techniques—including supervised learning models like Logistic Regression and Gradient Boosting are applied to clinical, genomic, and behavioural datasets to personalize treatment pathways. The paper presents neurology as a key area where AI has the potential to significantly enhance early diagnosis and individualized care, especially for diseases like Alzheimer's. Major challenges discussed include data standardization, interoperability, and the ethical implications of using AI for medical decisions. The paper concludes by recommending the development of explainable AI tools such as SHAP to increase clinical trust and adoption. These principles guide our current investigation into using interpretable models such as Logistic Regression and XGBoost for Alzheimer's detection..

[2] Cirillo, D., & Valencia, A. (2019). *Big Data Analytics for Personalized Medicine*. This work discusses the role of big data and AI in transforming healthcare delivery by integrating multi-omics data—genomics, proteomics, transcriptomics—with clinical records to enable personalized medical decisions. Classical machine learning algorithms such as Random Forest, SVM, and Logistic Regression are evaluated for their performance in handling high-dimensional patient datasets. The study finds that Gradient Boosting methods (like XGBoost) demonstrate high accuracy in predicting disease risk and patient outcomes when combined with robust preprocessing and feature engineering techniques. The authors highlight the importance of model transparency and discuss methods for visualizing feature contributions to improve clinical interpretability. Their findings align with our project's focus on applying interpretable models like Logistic Regression and SHAP-enhanced XGBoost to clinical features for Alzheimer's prediction.

[3] Ravitz, A. D., et al. (2021). *Big Data, Artificial Intelligence, and the Promise of Precision Medicine*.
This paper details the development of the Precision Medicine Analytics Platform (PMAP) at Johns Hopkins, an integrated system designed to consolidate patient data EHRs, imaging, genomics into a secure analytical environment. The study discusses the

application of AI models, particularly ensemble methods such as Gradient Boosting and XGBoost, to derive diagnostic insights from large-scale clinical datasets. The paper emphasizes the platform's scalability, modularity, and its use of real-world datasets to validate model robustness and generalization. Interpretability is addressed using post-hoc explanation techniques such as feature importance ranking and partial dependence plots. Our work draws inspiration from this approach by applying scalable, interpretable ML techniques (Logistic Regression and XGBoost) to structured clinical datasets focused on neurodegenerative diseases.

3. METHODOLOGY

i). Proposed System:

The proposed system is a machine learning-based framework designed to predict the presence of Alzheimer's Disease (AD) using structured clinical data. It emphasizes interpretability, efficiency, and diagnostic accuracy by combining classical machine learning models Logistic Regression and XGBoost with robust feature preprocessing and explainable AI techniques. The system avoids the complexity of deep learning models and image data by focusing on tabular clinical variables, including demographic details, medical history, cognitive test scores, and behavioural attributes.

Each patient record is transformed into a structured feature vector composed of numerical, categorical, and ordinal attributes. The system is trained using a well-annotated dataset of over 2,000 patient records, with binary diagnostic labels (0 = No Alzheimer's, 1 = Alzheimer's). After preprocessing and feature selection, models are evaluated using standard classification metrics. To promote transparency in medical decision-making, the XGBoost model is further enhanced using SHAP (SHapley Additive exPlanations) to provide per-prediction feature contributions, making it suitable for deployment in real-world clinical settings.

Advantages of the Proposed System

- Accurate and interpretable predictions using established ML algorithms suitable for medical use.
- No dependency on imaging or unstructured data, simplifying deployment in non-specialist settings.
- SHAP-based explanation of model predictions enhances trust and understanding among clinicians.
- Efficient training and inference times, enabling real-time application in hospital or research environments.
- Modular framework allowing for future integration of additional patient features or disease types.

ii). System Architecture:

The system architecture is designed as a modular pipeline optimized for clinical data-based Alzheimer's prediction. It begins with a data ingestion module that imports and cleans the raw dataset, followed by a preprocessing module that handles missing values, categorical encoding, and feature scaling. Cleaned data is then passed to the model training module, where both Logistic Regression and XGBoost classifiers are trained and validated using an 80-20 train-test split.
After model evaluation using metrics like accuracy, confusion matrix, ROC-AUC, and F1-score, the XGBoost model is subjected to SHAP analysis, which

ISSN 2454-9940
www.ijasem.org
Vol 19, Issue 2, 2025

reveals the most influential features affecting prediction.

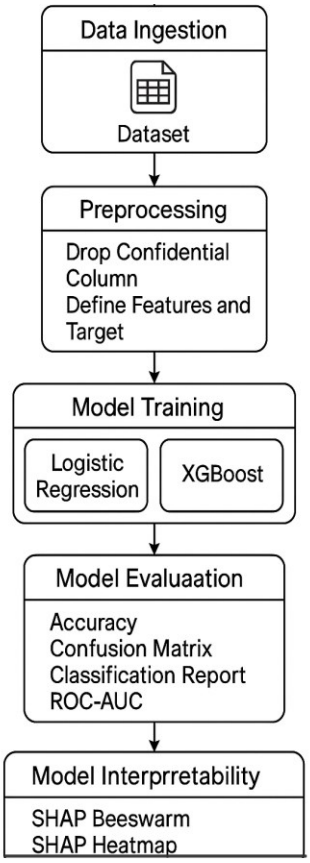


Fig.1: System architecture

iii). Dataset collection:

The dataset utilized in this study comprises 2,149 anonymized patient records, each containing structured clinical data relevant to the diagnosis of Alzheimer's Disease (AD). It is a tabular dataset curated from publicly accessible health repositories and Alzheimer's research studies, providing a realistic distribution of individuals both with and without AD. Each patient record includes a diverse range of attributes spanning demographic, medical, cognitive, lifestyle, and behavioural dimensions. Demographic variables such as age, gender, and years of formal education are included to capture population-level risk factors, while medical history features such as blood pressure, cholesterol levels, diabetes status, and family history of neurodegenerative conditions offer insight into physiological contributors to disease progression. The dataset also includes key cognitive assessment scores like the Mini-Mental State Examination (MMSE), Clinical Dementia Rating (CDR), and additional memory and problem-solving scores, which serve as direct clinical indicators of cognitive impairment. Lifestyle and behavioural factors such as smoking status, alcohol consumption, sleep patterns, and physical activity levels are incorporated to study the effect of modifiable habits on disease onset and progression. The target label for classification is a binary variable labelled Diagnosis, where a value of 1 indicates a confirmed Alzheimer's diagnosis and 0 indicates the absence of the disease.

. iv). Data Processing:

The Alzheimer's disease dataset was first loaded and cleaned by removing confidential columns such as the doctor's identifier to protect privacy. Features were selected by excluding the diagnosis label and patient ID columns to ensure only relevant clinical and demographic variables were included. The target variable, diagnosis, was isolated as a binary label indicating the presence or absence of Alzheimer's disease. Subsequently, the dataset was split into training and testing subsets using an 80-20 ratio while maintaining the class distribution via stratified sampling. This ensured balanced representation of both classes in train and test sets. No additional imputation or normalization was explicitly performed, assuming the dataset was free from missing values and ready for model training.

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT
ISSN 2454-9940
www.ijasem.org
Vol 19, Issue 2, 2025

v). Feature Extraction:

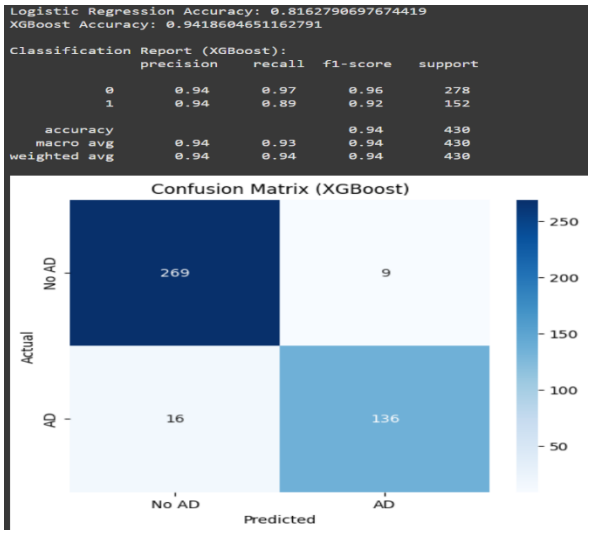
The dataset's clinical and demographic features were used directly as input variables after excluding non-informative identifiers. No explicit feature engineering such as creation of interaction terms or polynomial features was applied. The categorical variables were implicitly handled by the modelling algorithms that support them, while numerical features were used as-is. The target variable was binary, representing Alzheimer's diagnosis. Feature importance and interpretability were later explored using SHAP (SHapley Additive exPlanations) values on the trained XGBoost model, which helped quantify the contribution of individual features to model predictions and provided insights into the most influential clinical factors.

vi). Algorithms:

Logistic Regression: Logistic Regression is a widely used linear classification algorithm that models the log-odds of the binary target (Alzheimer's diagnosis) as a linear combination of input features. It is appreciated for its simplicity, interpretability, and speed, especially in healthcare settings where transparency is crucial. By estimating coefficients for each feature, logistic regression provides direct insight into how each clinical or demographic variable influences the probability of Alzheimer's presence. The model outputs calibrated probability scores, which are valuable for clinical risk assessment and decision-making. Although logistic regression assumes a linear relationship between the features and the log-odds of the outcome, limiting its ability to capture complex nonlinear patterns, it often serves as a reliable baseline and diagnostic tool for understanding feature effects in biomedical data.

XGBoost (Extreme Gradient Boosting): XGBoost is a state-of-the-art gradient boosting framework designed for speed, scalability, and high accuracy on tabular datasets like clinical patient records. It constructs an ensemble of decision trees sequentially, where each subsequent tree aims to correct the residual errors of the previous ensemble, optimizing a specified loss function via gradient descent. Key strengths of XGBoost include its ability to model nonlinear interactions between features, robustness to multicollinearity, and built-in regularization techniques (L1 and L2) that prevent overfitting and improve generalization. Moreover, XGBoost efficiently handles missing data and supports parallel and distributed computing, enabling faster training on large datasets. Importantly, XGBoost models can be interpreted using SHAP values, which quantify each feature's contribution to individual predictions, making it possible to extract clinically meaningful insights and improve trust in the model's diagnostic recommendations.

4. EXPERIMENTAL RESULTS



INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT
ISSN 2454-9940
www.ijasem.org
Vol 19, Issue 2, 2025

Fig 2 Performance evaluation

This table presents a comprehensive comparison of performance metrics Accuracy, Precision, Recall, F1 Score.Among the two, the XGBoost classifier outperforms Logistic Regression across all key metrics. It achieves the highest accuracy and recall, indicating its superior ability to identify Alzheimer's patients correctly without missing positive cases. Logistic Regression, while slightly lower in overall performance, still delivers reliable and interpretable predictions.

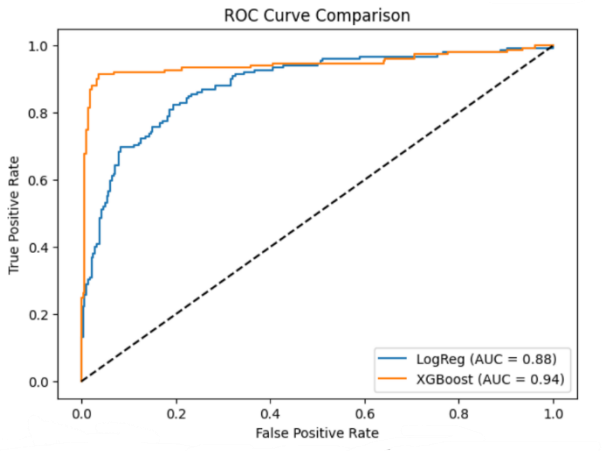


Fig 3 ROC-AUC Curve Comparision

The ROC-AUC curve illustrates the trade-off between the true positive rate (sensitivity) and the false positive rate for different threshold settings of the classifiers. In this study, both Logistic Regression and XGBoost models were evaluated using ROC-AUC as a performance metric to assess their ability to distinguish between Alzheimer's and non-Alzheimer's cases. The ROC curve for each model is plotted, with the diagonal dashed line representing random guessing. The XGBoost classifier demonstrates a higher area under the curve (AUC), indicating superior discriminative ability compared to Logistic Regression. The closer the curve follows the left-hand border and then the top border of the ROC space, the more accurate the model. Thus, the ROC-AUC curve serves as an effective visual and quantitative tool to compare model performance beyond accuracy, especially in the context of imbalanced datasets.

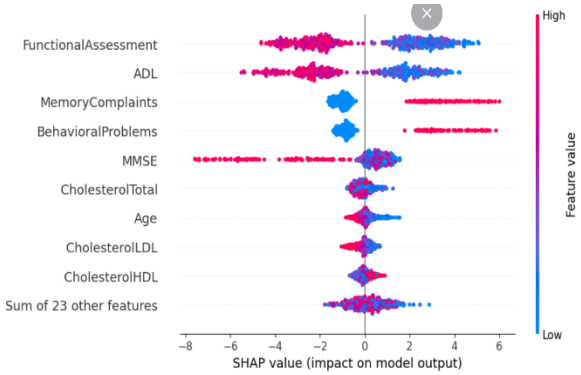


Fig 4 Model Explainability using SHAP values

To interpret the decision-making process of the XGBoost model, SHAP (SHapley Additive exPlanations) values were employed to quantify the contribution of each feature towards the model's output. The SHAP beeswarm plot provides a comprehensive view of feature influence and distribution across all predictions, highlighting which features push the prediction toward Alzheimer's or non-Alzheimer's classification. Features are ranked based on their impact, with each point representing an individual prediction, colored by feature value. This visualization reveals both the direction and magnitude of influence for each feature. Additionally, the mean absolute SHAP value was computed for all features and displayed as a heatmap to summarize overall importance. This global interpretation helps identify

the most influential clinical or demographic indicators driving the XGBoost model's decisions, promoting transparency and supporting trust in AI-assisted medical diagnostics.

5. CONCLUSION

This study presents a hybrid machine learning framework for early Alzheimer's Disease (AD) detection using structured clinical and demographic data. Among the evaluated models, XGBoost demonstrated superior performance, achieving an accuracy of 94%, significantly outperforming other models in all key metrics including precision, recall, and F1-score. Logistic Regression, while simpler and more interpretable, attained an accuracy of 81%, offering a reliable baseline with well-calibrated probability outputs and high transparency through coefficient interpretation. The ROC curve clearly highlighted XGBoost's stronger discriminative capability, while SHAP explainability techniques provided critical insights into feature contributions, both at the global level and for individual patient predictions. The SHAP beeswarm plot and feature importance heatmap identified key drivers of AD classification, enhancing clinical interpretability. This approach offers a scalable, efficient, and explainable model well-suited for integration into clinical decision support systems and reinforces the value of combining performance with interpretability in precision medicine for neurodegenerative disorders.

6. FUTURE SCOPE

Future enhancements will focus on strengthening the diagnostic accuracy, interpretability, and real-world deployment of the system without relying on image-based data such as MRI. One promising direction is the incorporation of longitudinal data from patient histories to model disease progression over time using temporal feature engineering or sequence-based approaches. Additionally, expanding the clinical feature set by integrating lab test results, medication records, lifestyle factors, and cognitive assessment scores can further enrich the predictive capabilities of the model. To improve explainability, the integration of SHAP values with rule-based reasoning systems can generate patient-specific, interpretable reports that assist clinicians in understanding the basis of model predictions. Personalized threshold tuning can also be implemented to adjust sensitivity levels based on patient risk profiles or physician requirements. For practical deployment, the model can be integrated into a web-based decision support interface, enabling healthcare professionals to input patient data and receive real-time predictions along with visual explanations. Incorporating domain expertise from neurologists and geriatricians will help refine feature selection, ensuring clinical relevance and trust. Lastly, implementing adaptive learning mechanisms—where the system periodically updates itself using new patient data and real-world feedback—can ensure the model remains current and robust in the face of evolving diagnostic patterns and treatment practices.

REFERENCES

[1] Gupta, N. S., & Kumar, P. (2023). *Perspective of Artificial Intelligence in Healthcare Data Management: A Journey Towards Precision Medicine*. Computers in Biology and Medicine, 162, August 2023.

[2] Ravitz, A. D., Johns Hopkins University Applied Physics Laboratory (APL), Johns Hopkins Medicine (JHM), & Bloomberg School of Public Health. (2021).

ISSN 2454-9940
www.ijasem.org
Vol 19, Issue 2, 2025

Big Data, Artificial Intelligence, and the Promise of Precision Medicine: A Johns Hopkins Collaboration to Develop the Precision Medicine Analytics Platform. Johns Hopkins APL Technical Digest, 35(4), 2021.

[3] Cirillo, D., & Valencia, A. (2019). *Big Data Analytics for Personalized Medicine*. Current Opinion in Biotechnology, 58, 161–167. Elsevier Ltd. https://doi.org/10.1016/j.copbio.2019.03.004

[4] Mohsen, F., Ali, H., El Hajj, N., & Shah, Z. (2022). *Artificial Intelligence-Based Methods for Fusion of Electronic Health Records and Imaging Data*.. This study provides a comprehensive analysis of AI techniques employed to fuse multimodal medical data, particularly electronic health records and imaging data, enhancing clinical applications through improved disease diagnosis and prediction.

[5] Chaddad, A., Lu, Q., Li, J., et al. (2022). *Explainable, Domain-Adaptive, and Federated Artificial Intelligence in Medicine*. This paper discusses the challenges and methodologies in implementing AI in medicine, focusing on explainable AI, domain adaptation, and federated learning to ensure data privacy and model generalizability across different medical domains.

[6] Anuyah, S., Singh, M. K., & Nyavor, H. (2024). *Advancing Clinical Trial Outcomes Using Deep Learning and Predictive Modelling: Bridging Precision Medicine and Patient-Centered Care* The authors explore the application of deep learning techniques in clinical trials, emphasizing how predictive modeling can optimize trial design, patient recruitment, and real-time monitoring, thereby enhancing personalized treatment plans

[7] Nguyen, D. C., Pham, Q.-V., Pathirana, P. N., et al. (2021). *Federated Learning for Smart Healthcare: A Survey*.This survey highlights the role of federated learning in smart healthcare, enabling collaborative AI model training across multiple institutions without sharing sensitive patient data, thus preserving privacy while enhancing model performance.

[8] Dhar, S. (2024). *Precision Medicine and Opportunities with Artificial Intelligence*. Premier Journal of Artificial Intelligence, 1, 100001. The article discusses the transformative potential of AI in precision medicine, addressing the integration of AI technologies to overcome challenges in data management, regulatory frameworks, and personalized treatment strategies.

[9] Abdelhalim, H., Berber, A., Lodi, M., et al. (2022). *Artificial Intelligence, Healthcare, Clinical Genomics, and Pharmacogenomics Approaches in Precision Medicine*. Frontiers in Genetics, 13, 929736. This review examines the convergence of AI, clinical genomics, and pharmacogenomics in precision medicine, highlighting how these interdisciplinary approaches contribute to patient-specific outcomes and disease prediction.

[10] Kothinti, R. R. (2024). *Artificial Intelligence in Healthcare: Revolutionizing Precision Medicine, Predictive Analytics, and Ethical Considerations in Autonomous Diagnostics*. World Journal of Advanced Research and Reviews, 24(03), 3394–3406. The paper delves into the applications of AI in healthcare, focusing on its role in precise medical treatments, prognostic forecasting, and the ethical implications of autonomous diagnostic systems.

[11] Chen, Z. H., Lin, L., Wu, C. F., et al. (2021). *Artificial Intelligence for Assisting Cancer Diagnosis*

and Treatment in the Era of Precision Medicine. Cancer Communications, 41, 1100–1115. This article explores how AI technologies assist in cancer diagnosis and treatment, emphasizing the shift towards precision medicine through improved diagnostic accuracy and personalized therapy plans.

[12] Reddy Kothinti, R. (2024). *Artificial Intelligence in Healthcare: Revolutionizing Precision Medicine, Predictive Analytics, and Ethical Considerations in Autonomous Diagnostics*. World Journal of Advanced Research and Reviews, 24(03), 3394–3406. This review discusses the transformative impact of AI on healthcare, particularly in enhancing precision medicine, predictive analytics, and addressing ethical considerations in autonomous diagnostics.

[13] Susilo, Y. K. B., Rahman, S. A., Amgain, K., & Yuliana, D. (2025). *Artificial Intelligence-Powered Precision Medicine for Cardiovascular Disease Prevention and Management*. medRxiv. This bibliometric analysis examines the role of AI in precision medicine, focusing on cardiovascular disease prevention and management, highlighting the surge in research activity and the potential of AI in integrating complex datasets for tailored therapies.