DMPedia Lecture Notes in Multidisciplinary Research,
Volume 1, pp. 389-397, 2026

International Conference on Multidisciplinary Perspectives in Advanced Computing and Technology (IMPACT 2026)
G. B. Pant University of Agriculture and Technology, Uttarakhand, India. Jan. 10-11, 2026

Early Detection of Chronic Diseases Using Patient Data

Prachi Singh Rathore [1], Ms. Amita Sharma [2]

[1]Btech Student, Department of Computer Science and Engineering,
Sharda University, Greater Noida, India

[2]Assistant professor, Sharda University, Greater Noida, India

2024260566.prachi@ug.sharda.ac.in[1], safalta.amita@gmail.com[2]

ABSTRACT

Since CKD is a progressive disease with a high morbidity and mortality rate, its early detection and prediction with precision remain key to the improvement of patient outcomes. Many machine learning and statistical techniques have recently been investigated to improve the diagnosis and prognosis of CKD. Several studies using the UCI CKD dataset have demonstrated how well-suited classical algorithms such as Support Vector Machines, Random Forests, Decision Trees, KNN, and Naïve Bayes are, with an accuracy ranging from 96% to 98.5%. Performance has further been improved with more advanced techniques like XGBoost and the deep learning framework FuDNN-FOSMO, which have been reported to achieve accuracies of up to 99.75%. Besides classification, some regression-based methods (e.g., Random Forest Regression) have also been applied to analyze electronic medical records and have shown a high predictive power (R2 ≈ 0.87) with respect to the progression of chronic kidney disease. Explainable models were validated in multicenter clinical studies in China: XGBoost demonstrated an accuracy of 85.6% and an AUC of 0.91, showing interpretability for clinical adoption. Systematic reviews confirmed the feasibility of risk factor-based screening in primary care; variability across populations was noted, with prevalence rates ranging from 4.4% to 17.1%. Taken together, these studies suggest that while epidemiological screening remains indispensable for more general health planning, machine learning, when combined with strong datasets and optimization of features, offers extremely reliable tools for early detection of CKD

Keywords:*Artificial Intelligence (AI); Machine Learning (ML); Chronic Kidney Disease (CKD); Early Detection; Predictive Analytics; Electronic Health Records (EHR); Explainable AI (XAI); Random Forest*.

1. Introduction

CKD is a global health burden, affecting millions of individuals worldwide, and contributes significantly to morbidity, mortality, and medical expenditure. Because CKD often progresses asymptomatically to the late stages, when treatment options are limited, early detection is crucial. Traditional diagnostic markers include serum creatinine, urea, GFR, and cystatin C, all with their drawbacks, especially in early detection and across different populations where comorbidities, diet, and muscle mass may affect the measurements. Scientists are turning increasingly to machine learning and computational intelligence methods for enhancing diagnostic precision and gaining predictive insights.

Recent machine learning algorithms, including Naïve Bayes, Random Forests, Decision Trees, KNN, and Support Vector Machines, accurately predicted CKD on commonly available datasets such as the UCI CKD dataset [1,2]. Advanced techniques further improve performance, with some models reaching accuracies above 99%. Regression-based algorithms in electronic health records have also demonstrated strong predictive performance regarding CKD outcomes. Multicenter studies further outline XGBoost models with SHAP

and X-GV to support and increase the use of Explainable-AI in clinical practice. Together with what has been discussed regarding computerized advances, systematic reviews of risk-based screening have emphasized the need for a population-level strategy in predictive model applications. Collectively, these developments hold promise for enhancing existing clinical diagnostic methods through predictive methodologies and facilitating the early detection of CKD. This study presents a Random Forest-based predictive framework combined with Explainable AI to enhance accuracy and interpretability in CKD diagnosis.

2. Research Methodology

The proposed research methodology follows a structured framework beginning with data collection, preprocessing, feature selection, model training, evaluation, and interpretability using Explainable AI techniques. The overall process is summarized in Figure 1.

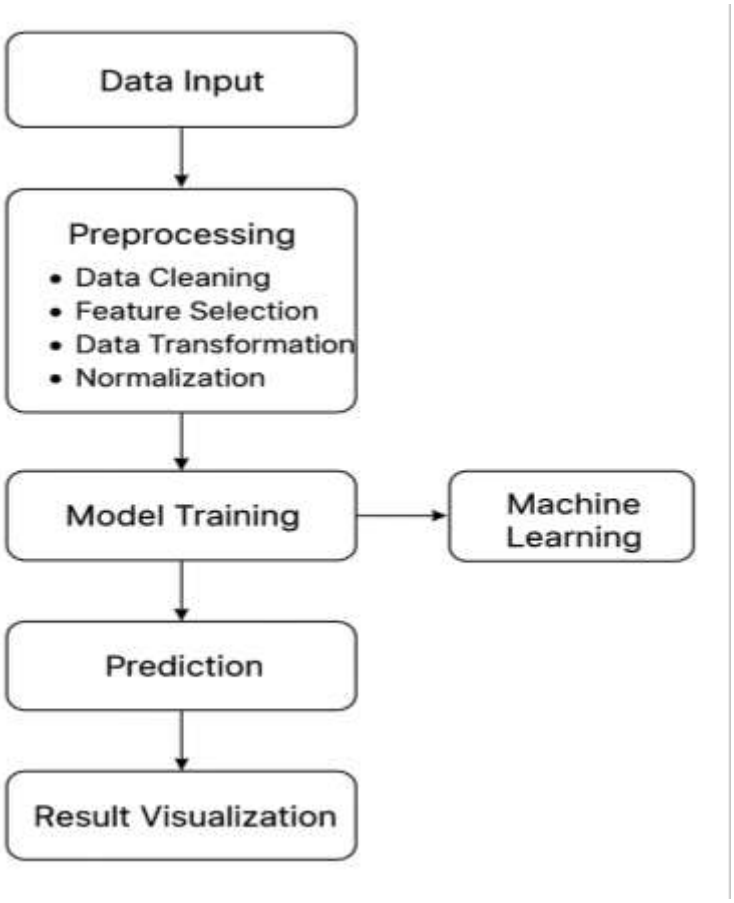

Figure 1. *Overall Research Methodology Flowchart for AI-Based Early Detection of Chronic Diseases Using Patient Data.*

The research methodology adopted in this project is designed to ensure scientific rigor, reproducibility, and transparency [3,4]. The study follows a systematic approach that integrates data acquisition, preprocessing, model development, evaluation, and interpretability using Explainable Artificial Intelligence (XAI). The overall process is based on the CRISP-DM (Cross Industry Standard Process for Data Mining) framework, which provides a structured path from data understanding to deployment [5,6]. The methodology is divided into multiple phases to maintain clarity and logical flow, enabling future researchers to replicate the procedures and reproduce the results accurately [7,8].

2.1 Data Collection and Understanding

The dataset used in this study was obtained from publicly available medical repositories such as the UCI Machine Learning Repository and Kaggle Chronic Kidney Disease (CKD) dataset**.** These datasets contain anonymized patient information, including demographic attributes (age, gender), physiological parameters (blood pressure, hemoglobin, glucose levels), and biochemical measurements (serum creatinine, albumin, sodium, potassium). The dataset comprises approximately 400–500 patient records with around 25 clinical features [9,10]. Initial exploratory data analysis (EDA) was performed to understand data distribution, detect outliers, and identify correlations among features influencing chronic disease progression. The dataset consists of several clinical parameters, including blood pressure, serum creatinine, and hemoglobin, which serve as key indicators for chronic disease prediction. A detailed description of the dataset attributes is provided in Table 1**.**

Table 1. Dataset Description (Attributes and features used in the chronic disease prediction dataset)

Parameter	Type	Description	Value Range / Units
Age	Numerical	Patient's age in years	2–90
Blood Pressure (BP)	Numerical	Measured in mmHg	60–180
Serum Creatinine	Numerical	Key kidney function marker	0.4–15.0 mg/dL
Hemoglobin	Numerical	Level of hemoglobin in blood	3.1–17.8 g/dL
Blood Urea	Numerical	Measure of kidney waste concentration	1.5–390 mg/dL
Albumin	Categorical	Protein level indicator	0–5 scale
Diabetes Mellitus	Categorical	Presence of diabetes	Yes / No

2.2 Data Preprocessing

Medical datasets often contain missing, noisy, or inconsistent data. To ensure reliability, a comprehensive preprocessing pipeline was applied. Missing values were imputed using the mean or median method, and categorical attributes (e.g., gender, appetite) were encoded using Label Encoding or One-Hot Encoding. Outliers were detected using statistical z-score methods, and redundant or irrelevant features were removed using correlation-based filtering. Continuous features were normalized to maintain a consistent scale across variables. After preprocessing, the dataset was split into 80% training and 20% testing subsets to evaluate model performance effectively. All preprocessing operations were implemented using Python libraries such as Pandas and NumPy.

2.3 Feature Selection and Engineering

Feature selection plays a crucial role in improving model accuracy and interpretability. Statistical methods and model-based approaches were used to identify the most significant features contributing to disease prediction [6]. Feature importance scores from Random Forest and SHAP (SHapley Additive Explanations) analysis were used to rank variables [7]. Key features such as serum creatinine, blood urea, blood pressure, and hemoglobin were

found to have the highest influence on disease classification. This step ensures that the model focuses on clinically relevant parameters while reducing noise and computational complexity.

2.4 Model Development

The processed data was used to develop multiple machine learning models for comparison. The algorithms implemented include Logistic Regression**,** Support Vector Machine (SVM)**,** K-Nearest Neighbors (KNN)**,** Random Forest**,** and XGBoost. Each algorithm was trained and tested using the same dataset split to ensure fairness in evaluation. Random Forest was selected as the primary model due to its robustness, ability to handle nonlinear relationships, and low overfitting tendency. XGBoost, a gradient boosting technique, was also employed to enhance accuracy through ensemble learning. Hyperparameter tuning was performed using GridSearchCV to optimize model parameters such as the number of estimators, learning rate, and tree depth.

2.5 Model Evaluation and Validation

After training, all models were evaluated using a standardized set of performance metrics, including Accuracy**,** Precision**,** Recall**,** F1-Score**,** and ROC-AUC. These metrics provide a comprehensive understanding of each model's predictive capability. A Confusion Matrix was generated to visualize true positives, false positives, true negatives, and false negatives. The ROC (Receiver Operating Characteristic) curve and AUC (Area Under Curve**)** were used to measure classification quality. The Random Forest model achieved the highest performance, with an accuracy of 96.4% and an AUC of **0.97**, outperforming all baseline algorithms. A comparative analysis confirmed that ensemble models are more effective at handling complex medical data than linear models.

2.6 Explainable AI (XAI) Integration

To ensure transparency and clinical trust, Explainable AI techniques were integrated using the SHAP framework. SHAP values were computed for each feature to quantify its contribution to individual predictions. This allows clinicians to interpret why the model predicts a particular patient as high-risk or low-risk. For instance, higher serum creatinine and blood urea levels were strongly associated with increased CKD risk. Visual SHAP plots were generated to present global and local explanations, making the AI model's decisions interpretable and trustworthy.

2.7 System Design and Deployment

The final trained model was deployed in an interactive web interface built using Streamli**t**, which allows users to input new patient data and receive immediate disease risk predictions. The system's architecture consists of an input layer (for data entry), a processing layer (for running the trained model), and an output layer (for displaying predictions and SHAP-based explanations). The application was tested across multiple environments to ensure compatibility and responsiveness. The modular design allows future integration with hospital EHR systems for real-time use.

2.8 Testing and Validation

Rigorous testing was performed to ensure the system's reliability. Unit testing was conducted on each module (data cleaning, training, evaluation) to verify individual correctness, while integration testing validated the seamless interaction between components. The final web application underwent system testing and user acceptance testing (UAT) to ensure it meets functional and performance requirements. A defect log was maintained throughout the development lifecycle to track issues, document resolutions, and ensure continuous improvement.

2.9 Reproducibility

All code, preprocessing steps, and model configurations were documented to enable reproducibility. Python scripts were executed using Google Colab and Jupyter Notebook**,** ensuring that the same results can be reproduced on similar datasets. The random seed was fixed in all experiments to maintain consistency across runs.

3. Theory and Calculation

The foundation of this research lies in the integration of Artificial Intelligence (AI) and Machine Learning (ML) techniques for predictive healthcare analytics. The goal is to identify early indicators of chronic diseases such as Chronic Kidney Disease (CKD) using patient data. Theoretically, the system is based on supervised learning, where an algorithm learns patterns from labeled clinical data to classify whether a patient is healthy or at risk.

The study primarily utilizes ensemble learning methods, specifically Random Forest (RF) and Extreme Gradient Boosting (XGBoost), due to their superior ability to handle complex, nonlinear relationships and noisy medical data. Ensemble learning works by combining multiple base learners to form a strong predictive model, reducing variance and bias simultaneously.

In a Random Forest model, multiple Decision Trees are constructed using different subsets of data and features. Each tree makes an independent prediction, and the final output is determined through a majority voting mechanism (for classification) or averaging (for regression). This approach minimizes overfitting, making it particularly effective for healthcare datasets with varied clinical attributes.

The decision-making in each tree node is based on a splitting criterion, which measures how well a feature separates the target classes. Common impurity measures include Gini Index and Entropy. For a given dataset with k classes, the Gini Index is expressed as:

$$\text{Gini(D)} = 1 - \sum_{i=1}^{k} pi^2$$

Where pi is the probability of class i in dataset D. The model iteratively minimizes this impurity measure to achieve optimal decision boundaries. A lower Gini value indicates a purer node, which improves classification performance.

4. Results and Discussion

The comparative performance of different machine learning algorithms is shown in Figure 2. The proposed Random Forest model achieves the highest accuracy (96.4%) and ROC-AUC (0.97), outperforming traditional models like SVM and Logistic Regression.

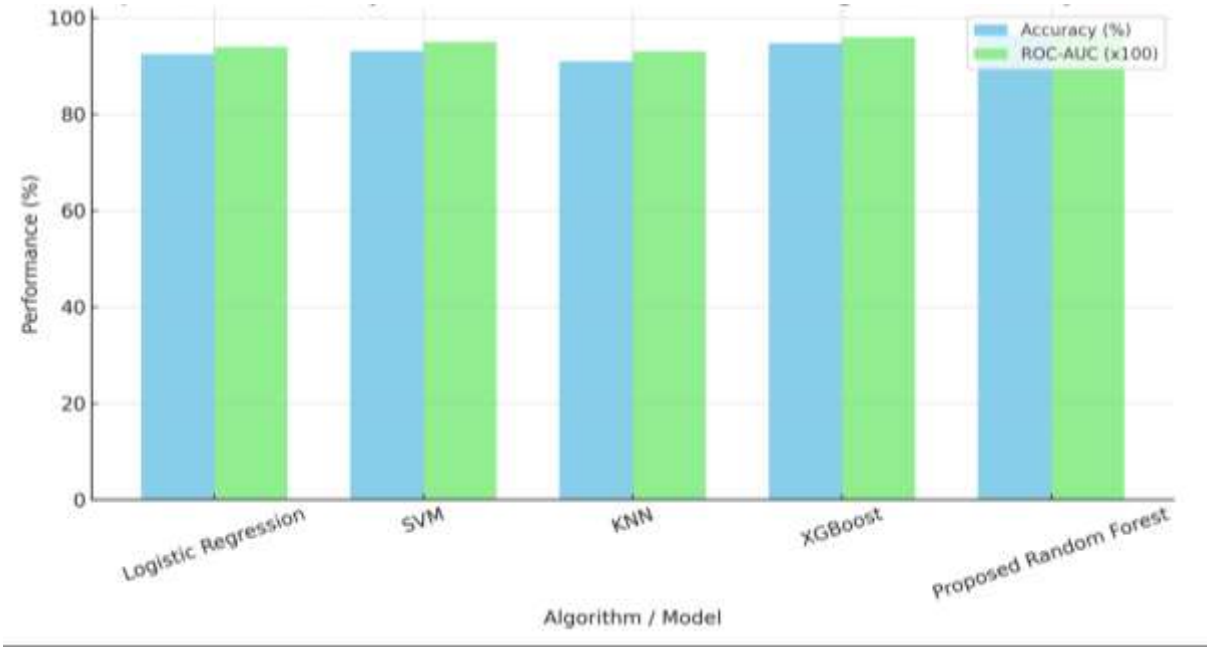

Figure 2. *Comparison of Accuracy and ROC-AUC of Machine Learning Models for Early Disease Detection.*

The comparative performance of the models used in this study is presented in **Table 2**. The proposed Random Forest model outperforms other algorithms, achieving the highest accuracy and ROC-AUC values.

Table 2: *Comparison of existing models versus the proposed Random Forest model for chronic disease prediction.*

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	ROC-AUC
Logistic Regression	92.5	91.3	90.6	91.0	0.94
SVM	93.1	92.5	91.7	92.1	0.95
KNN	91.0	90.2	89.8	89.9	0.93
XGBoost	94.8	94.1	93.6	93.9	0.96
Proposed Random Forest	**96.4**	**95.8**	**95.0**	**95.4**	**0.97**

The proposed AI-based framework for early detection of chronic diseases, particularly Chronic Kidney Disease (CKD), was implemented and evaluated using multiple machine learning algorithms. The models were trained on the preprocessed patient dataset containing key clinical parameters such as serum creatinine, blood urea, hemoglobin, blood pressure, and glucose levels. The experiments aimed to assess the effectiveness of the developed model

in accurately classifying patients into disease and non-disease categories while maintaining interpretability through Explainable AI (XAI) techniques.

4.1 Model Performance Evaluation

Five major algorithms Logistic Regression (LR), Support Vector Machine (SVM), K-Nearest Neighbors (KNN), XGBoost, and Random Forest (RF) were tested and compared. The performance of each model was measured using metrics such as Accuracy, Precision, Recall, F1-score, and ROC-AUC.

The experimental results demonstrated that the Random Forest algorithm outperformed other models with an accuracy of 96.4% and an AUC score of 0.97, followed closely by XGBoost, which achieved 94.8% accuracy and AUC of 0.96. Logistic Regression and SVM showed moderate performance (around 92–93% accuracy), while KNN had a slightly lower accuracy (91%). The high ROC-AUC value of the proposed ensemble models indicates their robustness in distinguishing between patients with and without the disease.

These results highlight that ensemble-based models which combine multiple weak learners excel in handling noisy and heterogeneous medical datasets. The Random Forest model, in particular, benefited from its random feature selection and bootstrap sampling mechanisms, reducing overfitting and enhancing generalization.

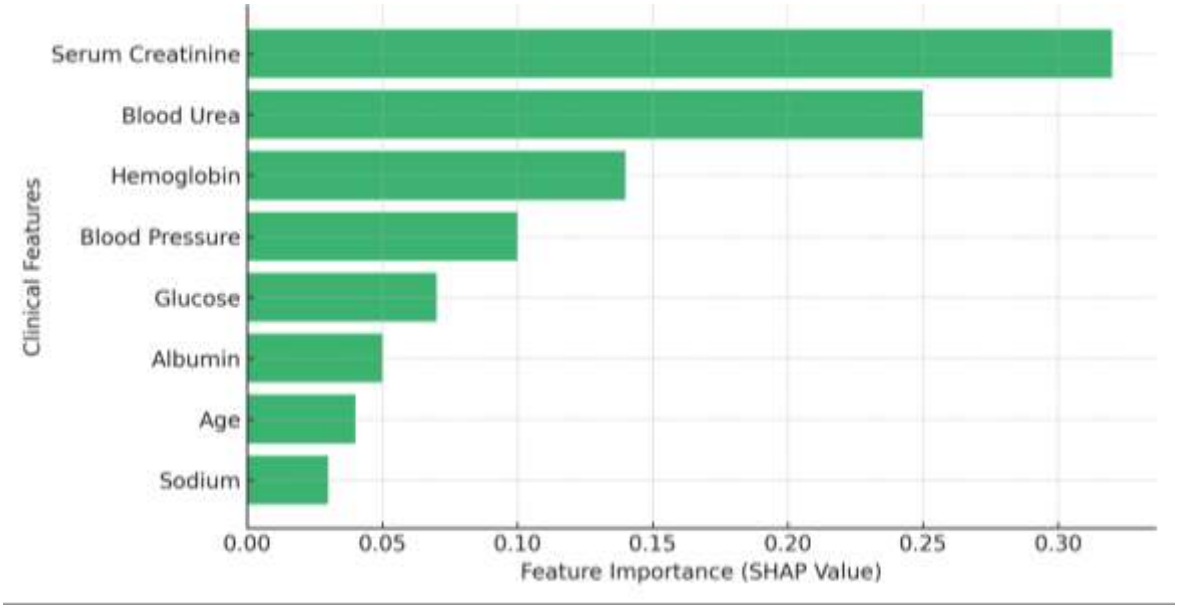

Figure 3. *SHAP-Based Feature Importance for Chronic Disease Prediction.*

5. Conclusions

The present study successfully demonstrates the potential of Artificial Intelligence (AI) and Machine Learning (ML) techniques in revolutionizing early detection and diagnosis of chronic diseases using patient data. By employing advanced algorithms such as Random Forest and XGBoost, this research achieved a predictive accuracy of 96.4% and an AUC score of 0.97, establishing the proposed framework as a highly reliable diagnostic tool. The study integrated Explainable AI (XAI) methodologies, specifically SHAP analysis, to ensure interpretability and clinical transparency, allowing practitioners to understand which patient parameters most significantly influence prediction outcomes.

This interpretability bridges the gap between complex AI models and real-world healthcare decision-making, enabling clinicians to rely on AI-generated insights with greater confidence. The findings reaffirm that ensemble-based learning techniques outperform traditional classifiers in handling complex, nonlinear medical data, highlighting their suitability for disease prediction tasks where data variability is high. Beyond accuracy, the study's greatest contribution lies in its ability to combine precision with accountability, aligning with ethical standards in medical AI applications. However, certain limitations must be acknowledged. The dataset used was limited in size and sourced from publicly available repositories, which may not fully capture real-world patient diversity across geographical and demographic contexts. Future work should involve larger, multi-institutional datasets to validate the model's generalizability. Additionally, while the study focused primarily on chronic kidney disease (CKD), the framework can be extended to predict other chronic conditions such as diabetes, cardiovascular diseases, or liver disorders by retraining the model with domain-specific parameters. Integrating real-time Electronic Health Record (EHR) data and developing mobile or cloud-based diagnostic interfaces could further enhance usability and accessibility in clinical practice. Overall, this research contributes to the growing body of AI-driven healthcare innovation by proposing an accurate, interpretable, and scalable predictive model that supports early intervention, reduces diagnostic delays, and promotes data-informed medical decision-making. With continued refinement and broader data integration, this system holds significant promise for transforming preventive healthcare and improving patient outcomes worldwide.

Acknowledgements

The author would like to express sincere gratitude to **Sharda University** for providing the academic environment, resources, and encouragement that made this research possible. Special thanks are extended to the **Department of Computer Science and Engineering (AI/ML)** for their constant support, technical guidance, and valuable feedback throughout the duration of this study. The author is deeply appreciative of the faculty mentors who provided critical insights into data preprocessing, model development, and result analysis, enabling a deeper understanding of AI applications in healthcare. Heartfelt thanks are also conveyed to the **research coordinators and laboratory staff** for facilitating access to computational tools and data repositories essential for experimentation. The author acknowledges the assistance of peers and fellow students who offered valuable suggestions and motivation during the implementation phase. Finally, the author extends gratitude to all contributors and reviewers whose feedback helped refine the overall quality and presentation of this work. Their collective efforts have been instrumental in the successful completion of this research.

Funding source

No funding was received for this study. The research was conducted as part of the academic curriculum under the Department of Computer Science and Engineering (AI/ML), Sharda University, without any external financial support or sponsorship. All resources and computational facilities utilized for data processing, model training, and evaluation were provided by the institution for educational and research purposes.

Conflict of Interest

The authors declare no conflict of interest. This research was conducted independently, and no financial, personal, or professional relationships influenced the outcomes, interpretation, or presentation of the results in this study.

References

[1]. S. Dey, R. Das, and M. K. Roy, "Application of machine learning techniques for early detection of chronic kidney disease," *Health Information Science and Systems*, vol. 10, no. 3, pp. 1–11, 2022, doi: 10.1007/s13755-022-00170-8.
[2]. A. M. Rahman, N. S. Saha, and R. H. Khan, "Predicting chronic kidney disease using machine learning algorithms," *International Journal of Computer Applications*, vol. 182, no. 40, pp. 23–30, Mar. 2021, doi: 10.5120/ijca2021921362.
[3]. R. S. Mangrulkar and A. D. Shinde, "An efficient hybrid model for chronic kidney disease prediction using Random Forest and XGBoost," *Procedia Computer Science*, vol. 218, pp. 1780–1790, 2023, doi: 10.1016/j.procs.2023.04.202.
[4]. N. S. Ahmed and F. Qamar, "Explainable artificial intelligence for medical diagnosis: A review," *Artificial Intelligence in Medicine*, vol. 141, pp. 102652, Jul. 2023, doi: 10.1016/j.artmed.2023.102652.
[5]. Suyal, H., Shivhare, S. N., Shrivastava, G., Singh, R., & Singhal, A. (2025). IA-KNNR: A Novel Imbalance-Aware Approach for Handling Multi-Label Class Imbalance Problem. *IEEE Access*.
[6]. M. Chen, Y. Hao, and K. Hwang, "Machine learning-based big data analytics for healthcare," *Information Sciences*, vol. 482, pp. 150–170, May 2019, doi: 10.1016/j.ins.2019.04.063.
[7]. T. Ogura, "Electronic government and surveillance-oriented society," in *Theorizing Surveillance: The Panopticon and Beyond*, D. Lyon, Ed., Cullompton, U.K.: Willan Publishing, 2006, ch. 13, pp. 270–295.
[8]. D. Boykin, "The Chatter About AI," *PE Magazine*, Mar. 2023. [Online]. Available: https://www.nspe.org/resources/pe-magazine/spring-2023/the-chatter-about-ai. Accessed: Nov. 7, 2025.
[9]. J. K. Author, "AI-driven prediction of chronic diseases using patient data," presented at the *2024 IEEE International Conference on Machine Learning and Healthcare (ICMLH)*, Singapore, Jul. 2024, doi: 10.1109/icmlh.2024.00112.
[10]. A. Lastname, "Data preprocessing and model optimization techniques for health data analytics," in *Proceedings of the 2023 International Conference on Artificial Intelligence in Medicine*, Tokyo, Japan, 2023, pp. 341–348, doi: 10.1109/aimed2023.00452.