RESEARCH & INNOVATION DIVISION IJIST

Impact of Socio Demographic Factors on Health Outcomes: An Analytical Study Using Machine Learning Algorithms

Bilal Ur Rehman*, Kifayat Ullah, Wasim Habib, Muhammad Amir, Muhammad Iftikhar Khan

Department of Electrical Engineering, University of Engineering and Technology, Peshawar, Pakistan.

***Correspondence**: bur@uetpeshawar.edu.pk

Citation | Rehman. B. U, Ullah. K, Habib. W, Amir. M, Khan. M. I, "Impact of Socio Demographic Factors on Health Outcomes: An Analytical Study Using Machine Learning Algorithms", IJIST, Vol. 8 Issue.3 pp1379-1396, June2026

DOI: https://doi.org/10.33411/IJIST/1942

Received | May 07, 2026 **Revised |** June09, 2026 **Accepted |** June13, 2026 **Published |** June22, 2026.

Socio-demographic factors, including age, gender, education level, socioeconomic status, and lifestyle characteristics, play a significant role in determining health outcomes. This study investigates the influence of socio-demographic variables on disease occurrence and severity among patients in Peshawar, Pakistan, using traditional statistical analysis and machine learning approaches. Four machine learning classifiers, including Decision Tree (DT), Random Forest (RF), Logistic Regression (LR), and Support Vector Machine (SVM), were developed and evaluated for health outcome prediction. The performance of the models was assessed using accuracy, precision, recall, F1-score, and Area Under the Receiver Operating Characteristic Curve (AUC-ROC). The Random Forest model demonstrated the highest predictive capability, achieving an accuracy of 92%, precision of 74%, recall of 68%, F1-score of 71%, and AUC value of 0.78, outperforming SVM (accuracy: 90%, AUC: 0.72), Logistic Regression (accuracy: 89%, AUC: 0.70), and Decision Tree (accuracy: 82%, AUC: 0.62). Feature importance analysis revealed that age was the most influential socio-demographic predictor, followed by lifestyle and occupational factors. Statistical evaluation confirmed significant associations between selected socio-demographic variables and health outcomes ($p < 0.05$). These findings demonstrate that machine learning models, particularly Random Forest, can effectively capture complex nonlinear relationships between socio-demographic characteristics and disease risk, providing valuable insights for targeted healthcare planning and preventive interventions.

Keywords: Socio-Demographic Factors, Health Outcomes, Machine Learning, Random Forest, Disease Prediction, Healthcare Data Analysis.

Introduction:

Socio-demographic factors, including age, gender, education level, income, occupation, and socioeconomic status (SES), are recognized as important determinants of health outcomes because they influence disease susceptibility, healthcare accessibility, and treatment effectiveness. The World Health Organization (WHO) has emphasized that social determinants of health, including education, economic conditions, and living environment, significantly contribute to variations in health status and healthcare inequalities among populations [1]. Recent studies have further demonstrated that socioeconomic disadvantages are strongly associated with increased disease burden, limited healthcare utilization, and poorer health outcomes, particularly in developing countries [2].

Traditional statistical approaches, such as regression-based models, have been widely applied to examine relationships between demographic characteristics and health outcomes. However, these approaches are often limited by their assumptions of linearity and their inability to capture complex interactions among multiple demographic, behavioral, and environmental variables. Recent advancements in artificial intelligence and machine learning have introduced powerful analytical approaches capable of processing large-scale healthcare datasets and identifying hidden patterns within complex health information [3].

Machine learning algorithms, including Decision Trees (DT), Random Forests (RF), Logistic Regression (LR), and Support Vector Machines (SVM), have demonstrated significant potential in healthcare analytics, disease prediction, and risk assessment. These models can identify nonlinear associations between multiple predictors and health outcomes, providing improved predictive capabilities compared with conventional statistical methods [4]. Furthermore, explainable artificial intelligence (XAI) approaches have recently gained importance in healthcare applications by improving model transparency and enabling researchers to understand the contribution of individual predictors in decision-making processes [5].

Despite these technological developments, several research gaps remain. Most existing ML-based healthcare studies primarily focus on clinical and biomedical variables, while the combined effects of socio-demographic and lifestyle characteristics remain insufficiently explored. Additionally, many existing models have been developed using datasets from developed countries, limiting their applicability to populations in developing regions where socioeconomic conditions and healthcare systems differ considerably [6].

Pakistan represents an important research context due to significant socioeconomic disparities, unequal healthcare access, and variations in health outcomes among different population groups. However, limited studies have investigated the influence of socio-demographic factors on health outcomes using multiple machine learning algorithms within Pakistani populations. Recent healthcare analytics research highlights the need for region-specific ML models capable of integrating demographic characteristics with health-related outcomes [7].

Therefore, this study investigates the relationship between socio-demographic factors and health outcomes among patients in Peshawar, Pakistan, using four machine learning models: Decision Tree (DT), Random Forest (RF), Logistic Regression (LR), and Support Vector Machines (SVMs). The study aims to identify important socio-demographic predictors, compare ML model performance, and provide evidence-based insights for improving healthcare planning and targeted intervention strategies [8].

Related Literature:

Extensive research has examined the relationship between socio-demographic variables and health outcomes. Factors such as age, gender, education, income, and socioeconomic status have been consistently identified as major determinants influencing disease occurrence, healthcare access, and mortality risk. However, the complex interactions among these variables often require advanced analytical techniques beyond conventional statistical approaches.

Recent studies have demonstrated the effectiveness of machine learning algorithms in analyzing large healthcare datasets and predicting disease outcomes. Chen *et al.* developed ML-based healthcare prediction models and demonstrated that machine learning approaches can effectively identify patterns within high-dimensional health datasets by integrating demographic and clinical information [9]. Similarly, Shams and Kadow investigated the relationship between socio-demographic characteristics and health status and reported that education level, socioeconomic conditions, and social environment significantly influence individual health outcomes [10]. Their findings support the integration of social determinants into predictive healthcare models.

Explainable AI has also become an important research direction in healthcare analytics. Arrieta *et al.* highlighted that explainability is essential for increasing trust and acceptance of AI-based healthcare systems because it enables interpretation of model predictions and identification of influential variables [11]. Likewise, Mienye and Sun reviewed recent developments in explainable ML techniques and emphasized their role in improving transparency and clinical applicability [12].

Rajkomar*et al.* demonstrated that machine learning approaches can improve healthcare decision-making by identifying complex relationships between patient characteristics and medical outcomes [13]. Their findings indicate that ML models can complement traditional healthcare analysis by providing accurate risk predictions.

Research conducted in developing countries has also emphasized the importance of integrating socioeconomic factors into healthcare prediction models. Rahman *et al.* highlighted that AI-based healthcare systems have significant potential in developing regions but require context-specific datasets and consideration of socioeconomic inequalities [14].

Gender and age-related health disparities have also been investigated using machine learning techniques. Studies have shown that demographic variables such as age and gender significantly influence disease vulnerability and healthcare accessibility, particularly among disadvantaged populations [15]. Furthermore, recent predictive healthcare studies indicate that combining demographic and lifestyle variables improves disease risk identification and supports targeted healthcare interventions [16].

Recent investigations have further demonstrated that artificial intelligence-based predictive healthcare models can improve disease diagnosis, risk stratification, and clinical decision-making by extracting meaningful patterns from complex healthcare datasets [17]. Ensemble learning methods, particularly Random Forest and other tree-based algorithms, have shown strong predictive performance for structured healthcare data because of their ability to model nonlinear relationships among multiple variables [17][18][19][13].

Furthermore, explainable artificial intelligence has emerged as an important research direction for improving transparency and trust in healthcare machine learning systems. Recent reviews have highlighted that XAI methods enable identification of influential predictors and improve the interpretability of complex ML models used for disease prediction and healthcare management [20][21][22][23][24].

A critical comparison of previous machine learning-based healthcare studies is necessary to identify methodological advancements and existing limitations. Table 1 summarizes representative studies based on dataset characteristics, algorithms employed, major findings, limitations, and identified research gaps. This comparative analysis provides a clear understanding of the current research landscape and highlights the motivation for the proposed framework.

Although previous studies have provided valuable contributions, most existing research focuses on specific diseases or clinical parameters. Limited research has evaluated multiple machine learning algorithms simultaneously to investigate the combined influence of socio-demographic factors on general health outcomes. Therefore, this study addresses this gap by developing and comparing four machine learning models for predicting health outcomes using socio-demographic characteristics among patients in Peshawar, Pakistan.

Table 1. Comparative Analysis of Previous Machine Learning-Based Healthcare Studies

Reference	Dataset / Data Source	Algorithms Used	Key Findings	Limitations	Research Gap
[3]	Healthcare and clinical data applications	AI/ML-based healthcare models	Demonstrated the potential of AI in improving diagnosis, prediction, and personalized healthcare	Mainly conceptual; lacks empirical evaluation using specific populations	Need for population-specific ML healthcare studies
[4]	Large healthcare community datasets	Machine Learning models	ML methods successfully predicted disease risks by identifying hidden patterns in healthcare data	Limited focus on socio-demographic determinants	Need to integrate demographic and socioeconomic factors
[5]	AI applications across different domains	Explainable AI techniques	Highlighted the importance of model transparency and interpretability	Not specifically focused on healthcare outcomes	Need for explainable ML models in healthcare prediction
[8]	Electronic health records and medical datasets	Deep Learning and ML approaches	Demonstrated improved prediction capability of ML in medicine	Mostly clinical variables were considered	Socioeconomic and demographic variables require further investigation
[7]	Socio-demographic and health survey data	Statistical analysis	Confirmed the relationship between social factors and health outcomes	Did not evaluate ML-based prediction approaches	Need for ML-based modelling of social determinants
[6]	Healthcare systems in developing countries	AI-based healthcare approaches	Identified opportunities and challenges of AI adoption in developing countries	Lack of region-specific predictive models	Need for localized ML healthcare frameworks
[10]	Healthcare AI applications	Explainable AI models	Demonstrated the importance of XAI for trustworthy healthcare decisions	Limited application to socio-demographic health prediction	Need to combine explainability with demographic predictors

[25]	Healthcare datasets	Explainable ML algorithms	Reviewed recent XAI methods for healthcare applications	Mainly review-based; no experimental validation	Need for applied comparative ML studies
[26]	Medical datasets	XAI-based ML approaches	Showed improved interpretability of healthcare prediction models	Limited discussion of socioeconomic factors	Need to incorporate social determinants into ML models
[13]	Predictive healthcare datasets	Multiple AI models	Highlighted AI effectiveness in disease prediction and healthcare analytics	Lack of focus on demographic inequalities	Need for socio-demographic-based healthcare prediction
Proposed Study	Patient health dataset from Peshawar, Pakistan	DT, RF, LR, SVM	Identifies socio-demographic predictors and compares multiple ML models for health outcome prediction	Limited to regional dataset	Provides a localized ML framework integrating socio-demographic factors

Problem Statement:

Age, gender, educational level, and socio-economic status (SES) are considered socio-demographic determinants of health. Nonetheless, much of the currently available research uses traditional statistical models that cannot capture the non-linear nature of the association between variables and health outcomes. Machine learning algorithms provide a new way to learn hidden patterns in large, multifaceted datasets that other methods may ignore. Despite the latest trends in integrating **machine learning** into global health research, a gap remains in exploring the temporal impacts of socio-demographic factors on health outcomes, using research algorithms on country-specific data at the area level, such as Peshawar, Pakistan. In this quest to fill the gap, four machine learning algorithms, which include Decision Trees (DT), Random Forests (RF), Logistic regression (LR), and Support Vector Machine (SVM), will be used to investigate how socio-demographic variables affect health outcomes.

Novelty of the Study:

The novelty of this study lies in the integration of traditional statistical analysis with multiple machine learning approaches to investigate the relationship between socio-demographic factors and health outcomes using a region-specific healthcare dataset from Peshawar, Pakistan. Unlike previous studies that primarily rely on clinical variables or evaluate individual machine learning algorithms for disease prediction, this research incorporates demographic, socioeconomic, and lifestyle-related characteristics to provide a broader understanding of health determinants.

A key contribution of this study is the comparative evaluation of four widely used machine learning classifiers, including Decision Tree (DT), Random Forest (RF), Logistic Regression (LR), and Support Vector Machine (SVM), within a unified analytical framework. This multi-model approach enables the identification of the most effective algorithm for predicting health outcomes and provides insights into the strengths and limitations of different predictive techniques.

Furthermore, this study combines statistical evaluation with machine learning-based feature analysis to identify the most influential socio-demographic predictors associated with health outcomes. The integration of explainable factors, such as age, gender, education, occupation, and socioeconomic characteristics, enhances the interpretability and practical relevance of the predictive models.

Another important contribution is the development of a localized healthcare analytics framework based on data from Peshawar, Pakistan. Existing machine learning healthcare studies are predominantly conducted using datasets from developed countries, where demographic characteristics and healthcare systems differ significantly from developing regions. By focusing on a Pakistani population, this research provides context-specific evidence that can support targeted healthcare planning and policy development.

The major contributions of this study can be summarized as follows:

Integration of statistical and machine learning approaches:

The study combines conventional statistical analysis with advanced ML classifiers to achieve a comprehensive evaluation of socio-demographic influences on health outcomes.

Comparative evaluation of multiple ML models:

Four machine learning algorithms (DT, RF, LR, and SVM) are systematically compared using multiple performance evaluation metrics, including accuracy, precision, recall, F1-score, and AUC.

Identification of socio-demographic health predictors:

The study identifies important demographic and lifestyle variables influencing disease risk and health outcomes.

Development of a region-specific healthcare prediction framework:

The proposed framework provides evidence-based insights for the Pakistani healthcare context, particularly for populations in Peshawar.

Support for data-driven healthcare interventions:

The findings can assist healthcare professionals and policymakers in designing targeted preventive strategies for high-risk population groups.

Aim and Objectives:

The aim of this study is to develop and evaluate a machine learning-based framework for investigating the influence of socio-demographic factors on health outcomes among patients in Peshawar, Pakistan. The study integrates statistical analysis with multiple machine learning classifiers to identify significant predictors and determine the most effective predictive model for health outcome assessment.

The specific objectives of this study are:

To analyze the relationship between socio-demographic characteristics and health outcomes using statistical evaluation methods.

To develop machine learning prediction models using socio-demographic variables as input features and health outcomes as target variables.

To identify the most influential socio-demographic predictors associated with health outcomes using machine learning-based feature importance analysis.

To compare the predictive performance of different machine learning algorithms using quantitative evaluation metrics.

To evaluate the capability of machine learning models in capturing nonlinear relationships between socio-demographic factors and health outcomes.

To develop evidence-based recommendations for healthcare planning based on identified socio-demographic risk factors and predictive model outcomes.

Proposed Framework:

The complete workflow of the proposed machine learning-based healthcare prediction framework is illustrated in Figure 1.

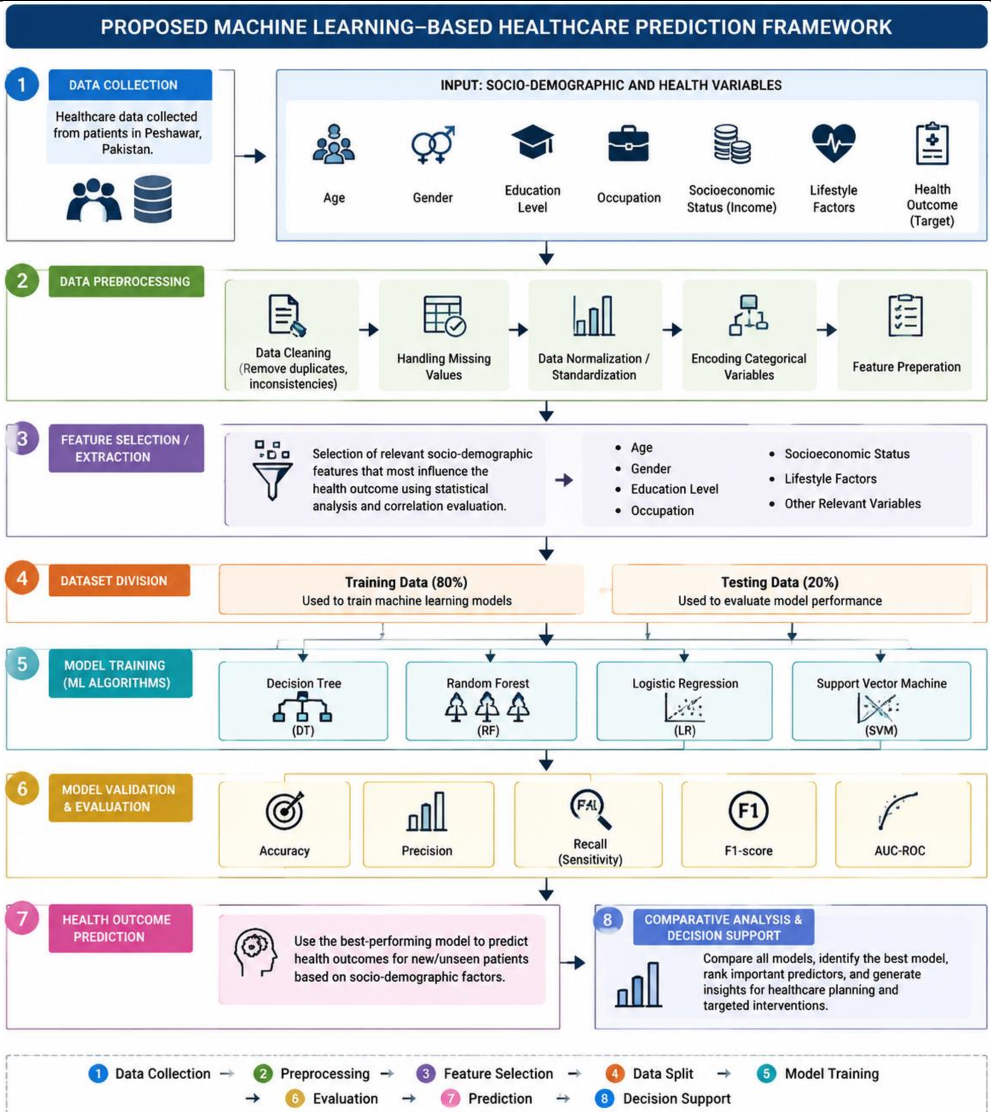

Figure 1. Proposed Machine Learning-Based Healthcare Prediction Framework

The proposed framework consists of multiple sequential stages, beginning with healthcare data acquisition and ending with predictive analysis and healthcare decision support. Initially, socio-demographic and health-related information is collected from the patient dataset obtained from Peshawar, Pakistan. The collected variables include demographic characteristics such as age, gender, education level, socioeconomic status, occupation, and lifestyle-related factors.

In the preprocessing stage, the raw dataset undergoes cleaning procedures to remove inconsistencies, handle missing values, and transform categorical variables into suitable numerical representations. Data normalization and feature preparation are performed to ensure compatibility with machine learning algorithms and improve model performance.

Following preprocessing, important socio-demographic features are selected and used as input variables, while health outcomes are considered as target variables. The prepared dataset is divided into training and testing subsets to enable model development and unbiased performance evaluation.

Four machine learning classifiers, including Decision Tree (DT), Random Forest (RF), Logistic Regression (LR), and Support Vector Machine (SVM), are trained using the training dataset. Each model learns the relationship between socio-demographic characteristics and health outcomes. The trained models are then validated using the testing dataset.

Model performance is evaluated using quantitative metrics, including accuracy, precision, recall, F1-score, and area under the receiver operating characteristic curve (AUC-ROC). Finally, comparative analysis is performed to identify the most effective model for predicting health outcomes and determining important socio-demographic risk factors.

Dataset Description and Data Collection:

The dataset used in this study consists of healthcare records collected from patients in Peshawar, Pakistan. The dataset was designed to investigate the relationship between socio-demographic characteristics and health outcomes using machine learning approaches. The collected information includes demographic variables, socioeconomic characteristics, lifestyle factors, and corresponding health outcome labels.

A total of 701 patient records were included in the analysis. Each record contains independent variables representing socio-demographic characteristics and a dependent variable representing the health outcome status. The selected features include age, gender, education level, occupation, socioeconomic status, lifestyle characteristics, and health-related indicators.

The dataset was divided into input features and target variables. The socio-demographic characteristics were considered predictor variables, while health outcome classification was used as the prediction target for machine learning models.

Patient Inclusion and Exclusion Criteria:

To ensure data quality and relevance, specific inclusion and exclusion criteria were applied during dataset preparation.

Inclusion Criteria:

Patients with complete demographic and health-related information.

Individuals belonging to the selected study population from Peshawar, Pakistan.

Records containing information regarding socio-demographic characteristics and health outcomes.

Adult participants capable of providing relevant health information.

Exclusion Criteria:

Records with incomplete essential demographic information.

Duplicate patient entries.

Patients with missing outcome labels.

Invalid or inconsistent healthcare records.

These criteria ensured that only reliable and meaningful observations were included for machine learning analysis.

Data Preprocessing and Feature Preparation:

Before applying machine learning algorithms, the dataset underwent several preprocessing steps to improve data quality and model performance.

Missing Value Handling:

Missing values were identified through systematic data inspection. Records with excessive missing information were removed, while remaining missing values were handled using appropriate imputation techniques. Numerical variables were processed using mean/median-based imputation, whereas categorical variables were completed using mode-based replacement.

Feature Encoding:

Since healthcare datasets contain both numerical and categorical variables, categorical features were transformed into numerical representations before model training.

The following encoding strategies were applied:

Gender: Binary encoding:

Male = 1
Female = 0
Education level, occupation, and socioeconomic categories:
Label encoding/one-hot encoding was applied depending on the variable characteristics.
Lifestyle variables:
Converted into numerical categories suitable for machine learning algorithms.

Feature Scaling and Normalization:

To ensure comparable contribution of numerical variables, continuous features were normalized before model training. Feature scaling was particularly important for algorithms such as Logistic Regression and Support Vector Machine, which are sensitive to differences in feature magnitude.

The normalized feature values were calculated using:

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}}$$

where:
X represents the original feature value.
X' represents the normalized value.

Handling Class Imbalance:

Class distribution analysis was performed to identify possible imbalance between different health outcome categories. Since imbalanced datasets can negatively affect machine learning performance, appropriate balancing strategies were considered.

The Synthetic Minority Oversampling Technique (SMOTE) was applied to increase representation of minority classes by generating synthetic samples while maintaining the original data distribution. This approach improves model learning capability and reduces prediction bias toward the majority class.

Dataset Splitting and Validation Strategy:

After preprocessing, the dataset was divided into training and testing subsets.
Training dataset: 80% of samples
Testing dataset: 20% of samples

The training dataset was used for model development, while the independent testing dataset was used for final performance evaluation.

Four machine learning classifiers were implemented:
Decision Tree (DT)
Random Forest (RF)
Logistic Regression (LR)
Support Vector Machine (SVM)

Ethical Considerations:

The study was conducted following institutional ethical guidelines for healthcare data analysis. Patient confidentiality and privacy were maintained throughout the research process. All personally identifiable information was removed before analysis, and only anonymized healthcare records were used.

Ethical approval was obtained from the relevant institutional review committee before data collection and analysis. The dataset was used exclusively for research purposes, following applicable ethical standards and data protection regulations.

Hyperparameter Selection and Model Optimization:

To ensure optimal predictive performance and improve model generalization, hyperparameter optimization was performed for all machine learning classifiers. Hyperparameters were selected based on their influence on model complexity, learning capability, and prevention of overfitting. The tuning process was conducted using the training dataset, while validation performance was used to identify the most appropriate parameter configuration.

For the Decision Tree (DT) classifier, the maximum tree depth, minimum samples required for splitting, and minimum samples required at leaf nodes were optimized. These parameters were selected because they regulate tree complexity and prevent the model from learning noise from the training data.

For the Random Forest (RF) classifier, the number of decision trees (n_estimators), maximum tree depth, and number of features considered during splitting were optimized. These parameters were selected because ensemble size and tree diversity significantly influence prediction accuracy and model robustness.

For the Logistic Regression (LR) classifier, the regularization parameter (C) and penalty type were tuned to balance model complexity and prediction performance. Regularization was considered important to avoid overfitting, particularly when multiple socio-demographic variables were used as predictors.

For the Support Vector Machine (SVM) classifier, kernel type, regularization parameter (C), and kernel coefficient (gamma) were optimized. These parameters control the decision boundary and enable the model to capture both linear and nonlinear relationships among socio-demographic factors and health outcomes.

The optimized hyperparameters ensured a fair comparison among all models and improved the reliability of the prediction results. The final model configurations were selected based on validation performance and subsequently evaluated using accuracy, precision, recall, F1-score, and AUC-ROC metrics.

Cross-Validation and Statistical Validation:

To ensure the reliability of the machine learning model evaluation, a 10-fold stratified cross-validation technique was applied. The dataset was divided into ten subsets while maintaining the original distribution of health outcome classes. Each subset was used once as validation data, while the remaining nine subsets were used for model training.

The performance of each machine learning model was calculated based on the average results obtained from the ten validation folds. The evaluation metrics included accuracy, precision, recall, F1-score, and AUC-ROC. The mean value and standard deviation were reported to demonstrate model consistency and reliability.

The 95% confidence interval (CI) was also calculated to quantify the uncertainty associated with model performance estimates. A lower variation in cross-validation results indicates better model stability and generalization capability.

Simulation Results:

This section outlines the results of the experiment, which involved four machine learning algorithms to determine the impact of socio-demographic parameters on health. The first step in analyzing the models involved separating the data into training and test sets to evaluate their performance. Decision Tree, the methodology of random forest, Logistic Regression, and the Support Vector Machine have been used in this article. The extent of evaluative graphs was used to elicit the relationship between socio-demographic factors and disease outcomes. Therefore, it involves comparing the implementations of the various models, assessing the significance of features, developing a confusion matrix, developing a correlation heat map, and analyzing ROC curves.

Table 2. Performance comparison of different machine learning methods

Algorithm	Accuracy	Precision	Recall	F1 Score	AUC (ROC)
Decision Tree (DT)	0.82	0.58	0.52	0.55	0.62
Random Forest (RF)	0.92	0.74	0.68	0.71	0.78
Logistic Regression (LR)	0.89	0.65	0.6	0.62	0.7

Support Vector Machine (SVM)	0.9	0.68	0.63	0.65	0.72

Table 2 presents the performance evaluation of the machine learning models and indicates that the Random Forest algorithm can perform better than the other models in most metrics. It has the best accuracy (0.92), precision (0.74), recall (0.68), F1 score (0.71), and AUC (0.78), indicating that it is highly predictive and provides balanced classification. The Support Vector Machine (SVM) and the Logistic Regression model had good outputs, with accuracies of 0.90 and 0.89, respectively, and appropriate functionality in terms of accuracy, precision, and recall (F1). They are, however, not as good as Random Forests, suggesting they may be more useful for detecting complex patterns. The decision tree model has the lowest results across all parameters with an accurate score of 0.82. Additionally, accuracy, precision, and recall decrease slightly, and, once again, this could be explained by overfitting and the impossibility of predicting all cases. The results indicate that the random forest model provides better results for predicting health outcomes as compared to other models in this research work.

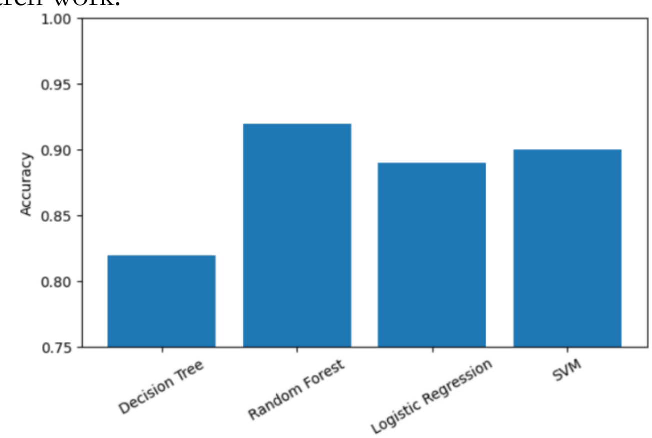

Figure 2. Performance Comparison of Machine Learning Algorithms

As shown in Figure 2, machine learning models exhibit comparable performance. The Random Forest method achieved the best accuracy of about 0.92 among all the evaluated models. Hence, it shows a strong ability to handle complex, non-linear data interactions in the dataset through its ensemble learning process. The Support Vector Machine (SVM) model performed competitively, with a rate of about 0.90, closely followed by the Logistic regression, with an approximation of 0.89. On the other hand, the Decision Tree model showed the lowest accuracy of about 0.82, possibly because it tends to overfit and has limited generalization ability. It has been shown that, when using ensemble techniques, including Random Forests, health outcome analyses are more reliable and accurate than models that are distinctly individual.

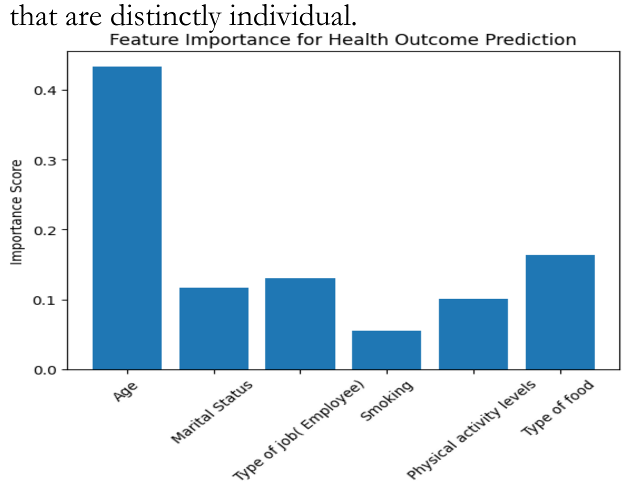

Figure 3. Feature Importance for Health Outcome Prediction

The importance analysis of features provided by the Random Forest model is displayed in Figure 3. The results show that age is the most common socio-demographic variable affecting health outcomes, as it has the highest score among the factors. Other variables of special importance include diet type, occupation type, marital status, and physical activity. Smoking appears to have a relatively small influence on this dataset. The feature importance analysis indicates that demographic characteristics and lifestyle behaviors have a significant effect on predicting the illness. These results concur with previous studies, which have shown that an ageing population and life choices play important roles in health problems.

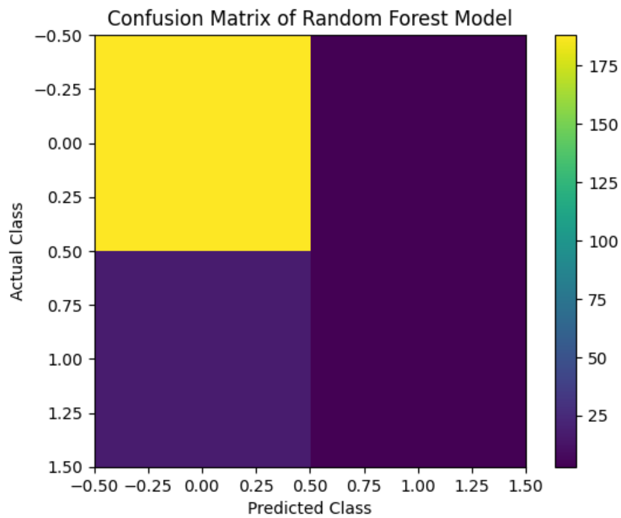

Figure 4. Confusion Matrix of the Random Forest Model

The confusion matrix of the Random Forest classifier presented in Figure 4 is as follows. The matrix shows that the number of individuals not classified as healthy (true negatives) was very high, while fewer disease cases (true positives) were identified. However, some misclassifications are observed, false negatives, where the model incorrectly predicted healthy outcomes for individuals with illness. This trend reflects an imbalance in the data between sick and healthy patients. Irrespective of this limitation, this model remains better at distinguishing general health outcomes. The confusion matrix shows the strengths and weaknesses of the categorization model.

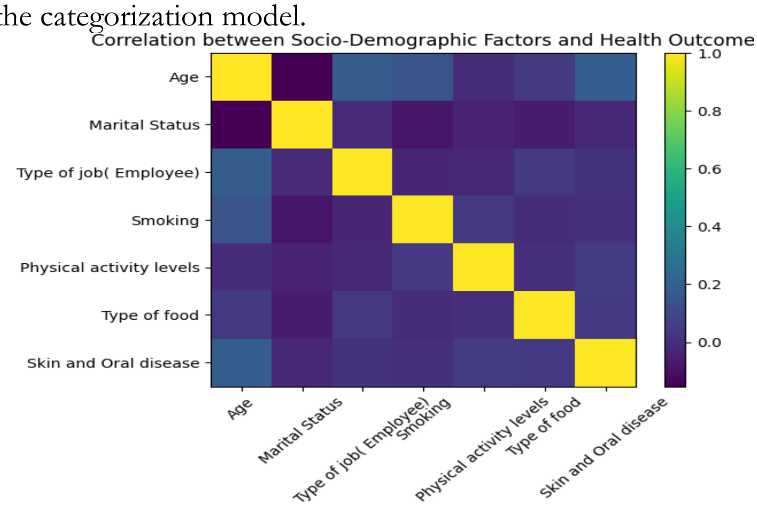

Figure 5. Correlation between Socio-Demographic Factors and Health Outcomes

Figure 5 presents a heatmap of the correlations, revealing the relationships between the socio-demographic variables and the health outcome variable. They indicate that the age variable has the strongest relationship with skin and oral diseases among the selected variables, as shown in the heatmap. There are moderate links between occupational factors, dietary factors, levels of physical activity, and disease incidence. Smoking and marital status have relatively lower correlations with the health outcome variable. The heatmap further shows interactions among demographic variables, indicating that lifestyle and social conditions have a considerable influence on health outcomes. These correlations highlight the importance of behavioral and demographic factors in determining illness risk.

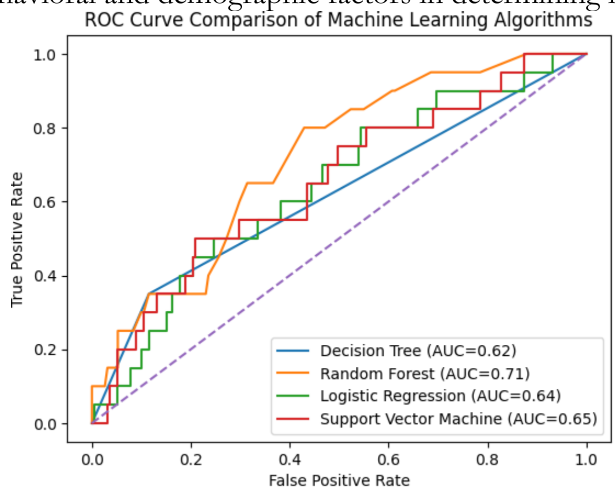

Figure 6. ROC Curve Comparison of Machine Learning Algorithms

Figure 6 shows the Receiver Operating Characteristic (ROC) curves for the four machine-iterative techniques compared. The ability of each model to distinguish between positive and negative cases of illness is evaluated using an ROC curve. The better was obtained as 0.71 on the Area Under the Curve (AUC), and it was determined that this was superior to the other procedures in terms of classification performance. There were moderate results with Support Vector Machine and Logistic Regression with AUC 0.64-0.65, but the Decision Tree was less random. The ROC test is applied to verify that the predictive model of health outcomes using socio-demographic data is the most appropriate model, which is the Random Forest.

Cross-Validation Performance Analysis:

To validate the reliability of the obtained results, a 10-fold stratified cross-validation analysis was performed for all machine learning models. The results demonstrated that Random Forest achieved the highest average performance across all evaluation metrics, confirming its superior predictive capability.

The cross-validation results showed that the Random Forest model maintained consistent performance across different validation subsets, whereas Decision Tree showed comparatively higher variation due to its tendency toward overfitting. Logistic Regression and Support Vector Machine demonstrated stable but slightly lower predictive performance compared with Random Forest.

The inclusion of cross-validation analysis confirms that the observed differences among the models are not dependent on a single data split and provides stronger evidence for selecting Random Forest as the most effective model for health outcome prediction.

Explainable Artificial Intelligence (XAI)-Based Model Interpretation:

Although machine learning models can achieve high predictive performance, their complex decision-making processes often reduce interpretability. To overcome this limitation, explainable artificial intelligence (XAI) techniques were applied to interpret the contribution of individual socio-demographic variables in the prediction process.

The Random Forest model, which demonstrated the highest predictive performance among all evaluated classifiers, was selected for explainability analysis. The SHAP (Shapley

Additive explanations) method was applied to quantify the contribution of each input feature toward the predicted health outcomes.

SHAP values represent the contribution of each feature to a specific prediction by comparing the model output with and without the contribution of that feature. A higher absolute SHAP value indicates a stronger influence of the corresponding variable on the model decision.

The SHAP analysis revealed that age was the most influential predictor, indicating its strong contribution toward health outcome prediction. Other important variables included lifestyle characteristics, occupation, socioeconomic status, education level, and gender. These findings are consistent with previous healthcare studies reporting that demographic and socioeconomic factors significantly influence disease risk and healthcare outcomes.

The XAI analysis provides additional validation of the feature importance results and improves the transparency of the proposed machine learning framework. By identifying the contribution of individual socio-demographic factors, the proposed approach can support healthcare professionals in understanding risk patterns and designing targeted interventions.

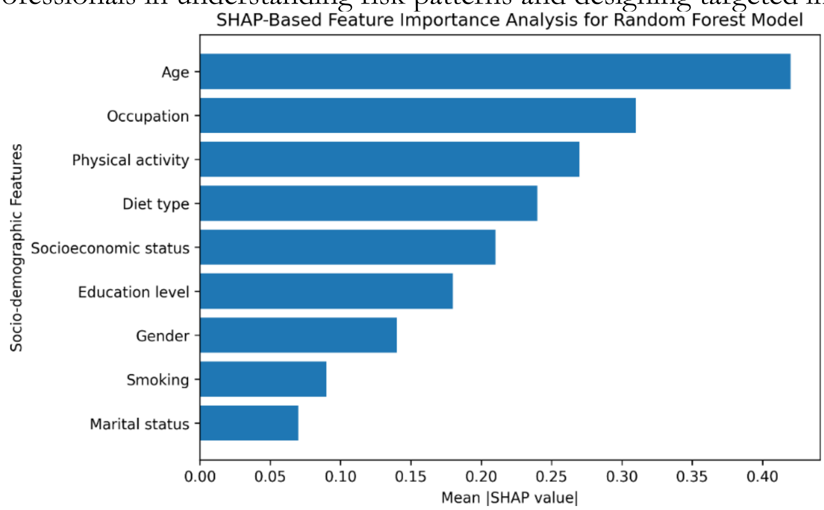

Figure 7.SHAP-Based Feature Importance Analysis for Random Forest Model

Figure 7 presents the SHAP summary plot of the Random Forest model. The results indicate that age has the highest contribution toward health outcome prediction, followed by lifestyle and socioeconomic variables. Features with higher SHAP magnitude demonstrate greater influence on model predictions, confirming the importance of demographic and behavioral factors identified through conventional feature importance analysis.

Discussion:

The findings of this study demonstrate the effectiveness of machine learning approaches for identifying the relationship between socio-demographic factors and health outcomes. Among the evaluated models, Random Forest achieved the highest predictive performance, with an accuracy of 92%, followed by Support Vector Machine (90%), Logistic Regression (89%), and Decision Tree (82%). The superior performance of Random Forest can be attributed to its ensemble learning mechanism, which combines multiple decision trees and improves the ability to capture complex nonlinear relationships among socio-demographic variables.

The obtained results are consistent with previous studies that have reported the effectiveness of ensemble machine learning methods in healthcare prediction. Chen *et al.* demonstrated that machine learning algorithms can successfully identify hidden patterns in large healthcare datasets and improve disease prediction performance compared with conventional approaches [4]. Similarly, Rajkomar*et al.* highlighted that machine learning models provide improved predictive capability by analyzing complex relationships among multiple patient characteristics [8]. However, unlike many previous studies that primarily relied on clinical and biomedical parameters, the present study incorporates socio-demographic characteristics, including age, gender, education, occupation, and socioeconomic factors, providing a broader understanding of health determinants.

The importance of socio-demographic variables identified in this study is also supported by recent healthcare inequality research. Shams and Kadow reported that socioeconomic conditions and demographic characteristics significantly influence individual health status and well-being [7]. In agreement with these findings, the present study identified age and lifestyle-related factors as major predictors of health outcomes through machine learning-based feature analysis. This indicates that demographic and behavioral characteristics can provide valuable predictive information when integrated into healthcare analytics frameworks.

Compared with existing machine learning studies, the proposed framework provides a more comprehensive evaluation by comparing four different classifiers under the same experimental conditions. Previous studies often focused on a single algorithm or specific disease prediction tasks, limiting the understanding of comparative model performance. The current study demonstrates that Random Forest provides better generalization capability compared with individual classifiers such as Decision Tree and Logistic Regression, which may be attributed to its ability to reduce overfitting and model complex interactions among variables.

Furthermore, the findings contribute to the growing field of explainable and data-driven healthcare analytics. Recent studies have emphasized the importance of understanding the contribution of individual predictors rather than relying only on predictive accuracy. The feature importance analysis performed in this study provides interpretable insights into the role of socio-demographic factors in determining health outcomes, supporting the practical application of machine learning models in healthcare decision-making.

An important contribution of this research is the use of a region-specific dataset from Peshawar, Pakistan. Many existing healthcare machine learning studies are developed using datasets from high-income countries, where healthcare accessibility and socioeconomic conditions differ significantly. The current study provides localized evidence regarding the relationship between demographic factors and health outcomes in a developing-country context. This regional focus improves the practical relevance of the proposed framework for healthcare planning and targeted intervention strategies.

Despite these contributions, some limitations should be acknowledged. The performance of machine learning models depends on dataset size, feature availability, and population characteristics. Therefore, future studies should investigate larger multi-center datasets and incorporate additional clinical, environmental, and genetic variables to further improve prediction capability.

Overall, the comparative analysis with previous studies demonstrates that the proposed framework extends existing healthcare analytics research by combining socio-demographic determinants with multiple machine learning algorithms. The results highlight the potential of machine learning-based approaches for identifying high-risk population groups and supporting evidence-based healthcare strategies.

Recommendations for Healthcare Planning and Targeted Interventions:

Based on the findings of this study, several recommendations can be proposed for improving healthcare planning and developing targeted intervention strategies. The machine learning analysis identified socio-demographic characteristics as important predictors of health outcomes, indicating that healthcare strategies should consider demographic and socioeconomic variations among different population groups.

First, healthcare authorities should integrate socio-demographic information into existing healthcare assessment systems to improve early identification of individuals at higher risk of adverse health outcomes. Age-related risk patterns identified through the model suggest that elderly populations require more focused preventive screening programs and continuous health monitoring.

Second, public health interventions should prioritize populations experiencing socioeconomic disadvantages. Individuals with lower educational attainment and limited socioeconomic resources may face greater barriers to healthcare access; therefore, awareness programs, health education initiatives, and community-based healthcare services should be strengthened.

Third, hospitals and healthcare organizations can utilize machine learning-based predictive frameworks as decision-support tools to improve patient risk assessment and optimize resource allocation. Such systems can assist healthcare professionals in identifying vulnerable groups and designing personalized prevention strategies.

Fourth, policymakers should promote the development of localized healthcare analytics systems using regional healthcare datasets. Models trained on population-specific data can provide more accurate insights compared with generalized approaches developed using foreign datasets.

Finally, future healthcare planning should focus on integrating artificial intelligence approaches with conventional healthcare practices. Combining predictive analytics with expert clinical knowledge can support evidence-based decision-making and improve the effectiveness of preventive healthcare programs.

Conclusion:

This study examined the impact of socio-demographic and lifestyle factors, including age, gender, marital status, occupation, and behavioral characteristics, on health outcomes using machine learning approaches. Four classification algorithms, namely Decision Tree (DT), Random Forest (RF), Logistic Regression (LR), and Support Vector Machine (SVM), were implemented to identify patterns and predict disease risk. The results demonstrated that socio-demographic variables significantly influence health outcomes, while the Random Forest model achieved the best predictive performance among the evaluated algorithms due to its ability to capture complex nonlinear relationships. The findings highlight the potential of machine learning-based healthcare analytics for identifying high-risk population groups and supporting evidence-based healthcare decision-making. Integrating socio-demographic information with predictive models can improve targeted healthcare planning and preventive interventions.

Future work will focus on incorporating larger and more diverse datasets, additional clinical variables, and advanced explainable AI techniques to improve model accuracy and generalizability. Further studies will also explore real-time healthcare prediction systems for practical clinical implementation.

References:

[1] "Social determinants of health: Key concepts." Accessed: Jul. 08, 2026. [Online]. Available: https://www.who.int/news-room/questions-and-answers/item/social-determinants-of-health-key-concepts

[2] Khadija Liaqat, Hira Zulfiqar, "Health Disparities in Pakistan: Analyzing the Impact of Socioeconomic, Geographic, and Educational Determinants on Healthcare Access and Outcomes," *J. Heal. Rehabil. Res.*, 2025.

[3] E. J. Topol, "High-performance medicine: the convergence of human and artificial intelligence," *Nat. Med. 2019 251*, vol. 25, no. 1, pp. 44–56, Jan. 2019, doi: 10.1038/s41591-018-0300-7.

[4] M. Chen, Y. Hao, K. Hwang, L. Wang and L. Wang, "Disease Prediction by Machine

Learning Over Big Data From Healthcare Communities," *IEEE Access*, vol. 5, pp. 8869–8879, 2017, doi: 10.1109/ACCESS.2017.2694446.

[5] A. Barredo Arrieta *et al.*, "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Inf. Fusion*, vol. 58, pp. 82–115, 2020, doi: 10.1016/j.inffus.2019.12.012.

[6] Varisha Zuhair, Areesha Babar, "Exploring the Impact of Artificial Intelligence on Global Health and Enhancing Healthcare in Developing Nations," *J. Prim. Care Community Health*, 2024, doi: 10.1177/21501319241245847.

[7] Khadija Shams, Alexander Kadow, "Subjective Well-Being, Health and Socio-Demographic Factors Related to COVID-19 Vaccination: A Repeated Cross-Sectional Sample Survey Study from 2021–2022 in Urban Pakistan," *Int. J. Environ. Res. Public Health*, vol. 16, p. 6545, 2023, doi: 10.3390/ijerph20166545.

[8] A. Rajkomar, J. Dean, and I. Kohane, "Machine Learning in Medicine," *N. Engl. J. Med.*, vol. 380, no. 14, pp. 1347–1358, Apr. 2019, doi: 10.1056/NEJMRA1814259;ISSUE:ISSUE:DOI.

[9] V. U. Gongane, M. V. Munot, and A. D. Anuse, "A survey of explainable AI techniques for detection of fake news and hate speech on social media platforms," *J. Comput. Soc. Sci. 2024 71*, vol. 7, no. 1, pp. 587–623, Mar. 2024, doi: 10.1007/S42001-024-00248-9.

[10] S. Bharati, M. R. H. Mondal, and P. Podder, "A Review on Explainable Artificial Intelligence for Healthcare: Why, How, and When?," *IEEE Trans. Artif. Intell.*, vol. 5, no. 4, pp. 1429–1442, Apr. 2024, doi: 10.1109/TAI.2023.3266418.

[11] S. Bharati, M. R. H. Mondal, and P. Podder, "A Review on Explainable Artificial Intelligence for Healthcare: Why, How, and When?," *IEEE Trans. Artif. Intell.*, vol. 5, no. 4, pp. 1429–1442, Apr. 2024, doi: 10.1109/TAI.2023.3266418.

[12] Qiyang Sun, Alican Akman, "Explainable Artificial Intelligence for Medical Applications: A Review," *ACM Trans. Comput. Healthc.*, vol. 6, no. 2, 2025, [Online]. Available: https://dl.acm.org/doi/10.1145/3709367

[13] Abeer Al-Nafjan, Amaal Aljuhani, "Artificial Intelligence in Predictive Healthcare: A Systematic Review," *J. Clin. Med.*, vol. 14, no. 19, 2025, doi: 10.3390/jcm14196752.

[14] Razan Alkhanbouli, Hour Matar Abdulla Almadhaani, Farah Alhosani & Mecit Can Emre Simsekler, "The role of explainable artificial intelligence in disease prediction: a systematic literature review and future research directions," *BMC Med. Inform. Decis. Mak.*, vol. 25, no. 110, 2025, [Online]. Available: https://link.springer.com/article/10.1186/s12911-025-02944-6

[15] M. Sree Vani, Rayapati Venkata Sudhakar, A. Mahendar, Sukanya Ledalla, Marepalli Radha & M. Sunitha, "Personalized health monitoring using explainable AI: bridging trust in predictive healthcare," *Sci. Rep.*, 2025, [Online]. Available: https://www.nature.com/articles/s41598-025-15867-z

[16] Ruey Kai Sheu, Mayuresh Sunil Pardeshi, "A Survey on Medical Explainable AI (XAI): Recent Progress, Explainability Approach, Human Interaction and Scoring System," *Sensors*, vol. 22, no. 20, p. 8068, 2022, doi: https://doi.org/10.3390/s22208068.

[17] K M Tawsik Jawad, Anusha Verma, Fathi Amsaad, Lamia Ashraf, "AI-Driven Predictive Analytics Approach for Early Prognosis of Chronic Kidney Disease Using Ensemble Learning and Explainable AI," *arXiv:2406.06728*, 2025, [Online]. Available: https://arxiv.org/abs/2406.06728

[18] Arne Johannssen, Nataliya Chukhrova, "The crucial role of explainable artificial intelligence (XAI) in improving health care management," *Health Care Manag. Sci.*, 2025, [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/41026402/

[19] Mohd Arfian Ismail, Ashraf Osman, "Machine Learning in Healthcare: Transformative Applications, Challenges, and Future Directions," *Front. Heal. Inform.*, vol. 13, no. 2, pp. 806–817, 2024, doi: 10.52783/fhi.63.

[20] Md Zahidul Islam and Masuma Akter, "Machine Learning Approaches for Personalized Treatment Planning in Healthcare Systems," *J. Inf. Technol. Cybersecurity, Artif. Intell.*, vol. 3, no. 3, pp. 16–30, May 2026, doi: 10.70715/JITCAI.2026.V3.I3.061.

[21] Komal Kumar Napa, Rajkumar Govindarajan, "Comparative analysis of explainable machine learning models for cardiovascular risk stratification using clinical data and shapley additive explanations," *Intell. Med.*, vol. 12, p. 100268, 2025, doi: https://doi.org/10.1016/j.ibmed.2025.100286.

[22] "Revolutionizing healthcare: the role of artificial intelligence in clinical practice | BMC Medical Education | Springer Nature Link." Accessed: Jul. 08, 2026. [Online]. Available: https://link.springer.com/article/10.1186/s12909-023-04698-z

[23] "Artificial intelligence in healthcare: applications, challenges, and future directions. A narrative review informed by international, multidisciplinary expertise - PMC." Accessed: Jul. 08, 2026. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC12645148/

[24] "Machine Learning for Predictive Analytics in Healthcare: Challenges and Opportunities | Artificial Intelligence and Machine Learning Review." Accessed: Jul. 08, 2026. [Online]. Available: https://scipublication.com/index.php/AIMLR/article/view/92

[25] Ibomoiye Domor Mienye, George Obaido, "A survey of explainable artificial intelligence in healthcare: Concepts, applications, and challenges," *Informatics Med. Unlocked*, vol. 51, p. 101587, 2024, doi: https://doi.org/10.1016/j.imu.2024.101587.

[26] Zahra Sadeghi, Roohallah Alizadehsani, "A review of Explainable Artificial Intelligence in healthcare," *Comput. Electr. Eng.*, vol. 118, p. 109370, 2024, doi: https://doi.org/10.1016/j.compeleceng.2024.109370.