African Journal of Science, Technology, Innovation and Development

ISSN: 2042-1338 (Print) 2042-1346 (Online) Journal homepage: www.tandfonline.com/journals/rajs20

Application of artificial intelligence in healthcare sector: A case study of predictive analytics model development for humanitarian healthcare system

Mawya Mourad, Olasukanmi Joseph Oladeji, Gift Funmilayo Gbadegeshin, Ousha Alshamsi, Taiwo Fimihan Olajumoke & Saheed Adebayo Gbadegeshin

To cite this article: Mawya Mourad, Olasukanmi Joseph Oladeji, Gift Funmilayo Gbadegeshin, Ousha Alshamsi, Taiwo Fimihan Olajumoke & Saheed Adebayo Gbadegeshin (03 Aug 2026): Application of artificial intelligence in healthcare sector: A case study of predictive analytics model development for humanitarian healthcare system, African Journal of Science, Technology, Innovation and Development, DOI: 10.1080/20421338.2026.2688442

To link to this article: https://doi.org/10.1080/20421338.2026.2688442

Published online: 03 Aug 2026.

African Journal of Science, Technology, Innovation and Development, 2026
https://doi.org/10.1080/20421338.2026.2688442

Routledge
Taylor & Francis Group

Application of artificial intelligence in healthcare sector: A case study of predictive analytics model development for humanitarian healthcare system

Mawya Mourad [1], Olasukanmi Joseph Oladeji [2,3], Gift Funmilayo Gbadegeshin [4], Ousha Alshamsi [5], Taiwo Fimihan Olajumoke [1,6] and Saheed Adebayo Gbadegeshin [1,7*]

[1]*Department of Supply Chain and Project Management, Arden University, United Kingdom*
[2]*Department of Medical Laboratories, Institut Universitaire Edexcel Du Benin (Edexcel University of Benin), Benin Republic*
[3]*His Favour Medical and Diagnostic Centre, Nigeria*
[4]*Department of Postgraduate Studies, University of Greater Manchester, United Kingdom*
[5]*Department of Engineering, Abu Dhabi University, United Arab Emirates*
[6]*Department of Postgraduate Studies, University of Worcester, United Kingdom*
[7]*College of Business and Economics, University of Johannesburg, South Africa*
**Corresponding author email: sgbadegeshin@arden.ac.uk; saheed.gbadegeshin@gmail.com*

Access to quality healthcare is a persistent challenge for refugee populations globally, particularly in conflict-affected regions with limited resources. The use of Artificial Intelligence (AI) might ease the challenge. Thus, this article explores the development of a predictive analytics model for forecasting patient outcomes within the healthcare system operated by the United Nations Relief and Works Agency for Palestine Refugees in the Near East (UNRWA) in Syria. A Machine Learning (ML) model development process was used to build a predictive analytics model capable of identifying patients at high risk of hospital readmission within 30 days of discharge, using synthetic fromdata with 1000 patients. The model was evaluated and achieved an accuracy of 88%, suggesting its potential effectiveness in supporting early intervention and resource prioritization in humanitarian health settings. Thus, the model can be a potential tool for data-driven decision-making that can improve healthcare outcomes within the UNRWA health system. To provide insights for scholars and practitioners, a framework titled 'Humanitarian ML Readiness Framework (H-MLRF)' is proposed. The framework suggests key activities and their deliverables for developing and applying predictive analytics models in humanitarian settings. Thus, the article contributes to discourse on development of predictive analytics and ML models.

Keywords: predictive analytics model, humanitarian healthcare system, artificial intelligence

Introduction

Predictive analytics has emerged as a transformative tool in modern healthcare, offering new opportunities to anticipate patient outcomes, enhance clinical decision-making, and optimize healthcare delivery. It refers to the use of statistical techniques, machine learning algorithms, and historical data to forecast future events or trends within a healthcare setting (Arowoogun et al. 2024). Unlike traditional retrospective analysis, predictive analytics enable proactive interventions by identifying patients at risk of adverse events such as hospital readmissions, disease progression, or treatment complications.

Over the past decade, healthcare systems in high-income countries have increasingly adopted predictive analytics to improve outcomes and reduce costs. One widely studied application is the prediction of hospital readmission risk, particularly for patients with chronic conditions such as heart failure, diabetes, and chronic obstructive pulmonary disease (Kansagara et al. 2011). By identifying patients at high risk of readmission before discharge, hospitals can implement targeted interventions, such as follow-up calls, medication adjustments, or personalized care plans, thereby reducing avoidable readmissions and improving patient safety. Another significant application lies in early warning systems, where predictive models are integrated into electronic health records to monitor patient vitals in real time and flag deteriorating conditions.

Predictive analytics has also been applied to population health management, enabling healthcare systems to stratify patients by risk and allocate preventive resources accordingly. Large health systems, such as Kaiser Permanente and the Veterans Health Administration in the United States, have successfully used predictive tools to guide chronic disease management, vaccination campaigns, and mental health support (Beam and Kohane 2018). Similarly, in the United Kingdom, the National Health Service (NHS) has incorporated such predictive tools to enhance early detection of clinical deterioration, using Modified Early Warning Score as key benchmarks. This kind of system enhances clinical responsiveness and supports data-informed decisions that can be lifesaving in high-pressure environments.

Despite the growing body of evidence supporting predictive analytics, its adoption remains uneven. Most success stories come from well-resourced environments with access to large, high-quality datasets, advanced infrastructure, and skilled personnel. In contrast, low-resource settings – especially those operating in humanitarian contexts – face considerable barriers to leveraging

African Journal of Science, Technology, Innovation and Development is co-published by NISC Pty (Ltd) and Informa Limited (trading as Taylor & Francis Group)

these technologies. These include limited digital health infrastructure, lack of standardization in data collection, and constraints in technical capacity. However, recent advances in open-source machine learning platforms have increased awareness of the benefits of data-driven healthcare, and the growing availability of anonymized datasets have started to shift this trend. Pilot projects in countries such as Rwanda and India have demonstrated that simplified, interpretable predictive analytics models can be deployed effectively even in low-resource environments, where they can contribute to maternal health monitoring, infectious disease surveillance, and triaging emergency cases (Davenport and Kalakota 2019).

Thus, this article aims to analyze roles of Artificial Intelligence (AI) in the healthcare sector by exploring how predictive analytics model, particularly through machine learning (ML) model, can be applied to improve patient outcomes. To achieve its aim, this article provides answers to the following research questions.

- How can predictive analytics be developed in order to forecast patient situations, such as hospital re-admission and complication risks?
- How can predictive analytics model be used to improve healthcare services in humanitarian environments?

These questions aim to examine how ML models can be designed and trained to analyze patient data, such as demographics, medical history, and vital signs to predict adverse health events before they occur. In high-resource settings, such models have already demonstrated success in forecasting hospital readmissions and complications, particularly when applied to structured clinical data (Rajkomar, Dean, and Kohane 2019). However, this article explores the feasibility and potential adaptation of such models in a low-resource, especially conflict-affected environment such as Sudan, Syria, and Gaza, where infrastructure, digital systems, and real-time data are often limited. In short, it investigates humanitarian healthcare systems. Its main goal is to understand how predictive analytics models can support proactive interventions, reduce preventable hospital visits, and ultimately improve patient care in this challenging context. The article is structured as follows: research background, research methodology, application of predictive analytics model, practical implications and conclusion.

Research background

The healthcare sector is among the most highly regulated industries because it concerns people's safety (Gbadegeshin, 2019b). Thus, its operations consist of systematic processes for the efficient, effective and safe delivery of medical services (McLaughlin, Olson, and Sharma 2022). Meanwhile, the advent of digitalization has helped shape the healthcare sector, especially its processes and management (Gbadegeshin 2019a, 2019b; Shaheen 2021). One of the digital technologies that shapes healthcare is Artificial Intelligence (AI), which is described as a system capable of acting independently with little or no human input (Gbadegeshin et al., 2021). AI has been applied in healthcare organizations for different purposes, such as improving accuracy and efficiency, promoting better experience of patients and reducing costs. Hence, it influences how various healthcare operations are conducted (Nichols 2025; Saraswat et al. 2022; Väänänen et al. 2021).

To improve accuracy and efficiency, many healthcare organizations are integrating AI into their clinical activities to improve their clinical decision-making by allowing AI to access real-time patient data. This enables clinicians to obtain up-to-date, relevant information about patients to make informed decisions. With the integration of AI, healthcare organizations are able to offer continuous support to their patients and ensure continuity of care (Väänänen et al. 2021). These are possible with the predictive capability of AI (Saraswat et al. 2022; Väänänen et al. 2021). This capability is achieved by predictive analytics, which depends on ML models.

ML is the ability of a computing model to identify complex, nonlinear relationships among variables and to generate accurate predictions from multidimensional datasets. Unlike traditional statistical methods that often assume linearity and normality, ML offers flexibility and adaptability, essential traits when dealing with the uncertainty and variability of healthcare data (Bzdok et al. 2018). While ML has rapidly gained traction in well-resourced healthcare systems, its application in low-resource and humanitarian settings is still emerging. These environments are typically characterized by limited infrastructure, fragmented or paper-based data systems, scarcity of trained personnel, and frequent disruptions due to conflict or displacement. However, the pressing healthcare needs in such contexts make them prime candidates for innovative, scalable, and cost-effective solutions, with predictive analytics as a promising example (Yozwiak et al. 2019).

A major barrier to implementing ML in humanitarian healthcare is the limited availability of real-time data and the absence of integrated digital systems. Clinics are operated by agencies such as the United Nations High Commissioner for Refugees (UNHCR), World Health Organisation (WHO), and United Nations Relief and Works Agency (UNRWA). These agencies often function with minimal electronic health record (EHR) systems, if any. Consequently, the ability to gather sufficient, structured data for training ML models is limited (Wirtz et al. 2017). Nevertheless, recent research has shown that synthetic data generation, transfer learning, and lightweight ML algorithms can help bridge some of these gaps, enabling meaningful predictions even with small or incomplete datasets (Fritz et al. 2021).

In field studies, ML has been successfully used to support disease surveillance, triage systems, and resource forecasting in fragile settings. For example, during the Ebola outbreak in West Africa, predictive analytics models were deployed to forecast disease spread and identify high-risk communities, enabling more focused interventions (Fleming 2017). In South Asia and Sub-Saharan Africa, ML models have been implemented to predict maternal mortality, malnutrition risk, and vaccine hesitancy with reasonably high accuracy, despite data scarcity and infrastructure limitations

(Saria, Butte, and Sheikh 2018). The key to success in these settings is adaptability. Unlike high-resource contexts where computational complexity and data size are not limiting factors, models designed for humanitarian use must prioritize interpretability, ease of use, and minimal computational demands.

Despite growing interest in applying ML to healthcare, there remains a significant gap in both the academic literature and practical applications regarding its use in humanitarian settings. Most existing studies and predictive healthcare models have been developed and tested in high-resource environments, where access to large, structured datasets, robust digital infrastructure, and highly skilled personnel is assumed (Beam and Kohane 2018; Kansagara et al. 2011). As a result, the existing literature often overlooks the constraints and ethical sensitivities that are unique to low-resource, crisis-affected healthcare environments. Furthermore, the existing ML models often require real-world patient data for training, something that is difficult or unethical to access in 'refugee contexts' due to privacy, security, and political concerns (International Committee of the Red Cross, ICRC 2020; UNHCR 2021). Hence, the novelty of this article lies in its intersection of ML and humanitarian health services. Its research process is discussed in the next section.

Methodology

This article adopted a machine learning (ML) model development process involving the systematic transformation of raw data into a deployable predictive system. The process typically begins with problem definition, in which objectives, success criteria, and operational constraints are specified clearly (Bishop 2006). It is followed by data collection and preprocessing, including data cleaning, handling missing values, normalization, and feature engineering to improve model performance (Géron 2019). The dataset is then divided into training, validation, and test sets to support generalizability. The process also includes model selection, whereby appropriate algorithms, such as decision trees, support vector machines, or neural networks, are chosen in line with the problem type and data characteristics (Murphy 2012). During training, the model learns patterns by minimizing a loss function through optimization techniques such as gradient descent. Hyperparameter tuning, often implemented through cross-validation, is then used to improve predictive accuracy and reduce overfitting. The process concludes with model evaluation, which relies on metrics such as accuracy, precision, recall, F1-score, ROC-AUC (receiver operating characteristic area under the curve), or mean squared error, depending on whether the task is classification or regression (James et al. 2021). Once validated, the model can be deployed into practice, while continuous monitoring and retraining help to prevent model drift and maintain long-term reliability.

Given the aim and research questions of this article, the ML model development process was appropriate because it enabled a structured exploration of predictive healthcare modelling (Chen et al. 2021). The article is exploratory and developmental in nature, seeking not only to produce a technically sound predictive model but also to assess the feasibility, value, and limitations of deploying predictive analytics within the operational realities of low-resource healthcare systems. For this reason, the UNRWA refugee healthcare context in Syria was selected as the empirical setting. UNRWA is the primary provider of healthcare services for Palestinian refugees across the Middle East, including Syria. It provides a comprehensive package of care comprising primary healthcare, maternal and child health services, chronic disease management, and psychosocial support for more than five million refugees. However, its operations are constrained by underfunding, supply shortages, damaged infrastructure, and limitations in digital data systems (UNRWA 2023). In such a setting, innovative digital solutions are needed to improve healthcare delivery and planning. Predictive analytics can help shift health systems from reactive to proactive approaches by identifying at-risk patients, optimizing resource allocation, and enabling earlier interventions. In addition, the protracted civil war in Syria, ongoing since 2011, has led to the destruction of healthcare infrastructure, repeated displacement, and widespread poverty. As a result, access to consistent and high-quality healthcare has become increasingly difficult for refugee communities (Smith, Jones, and Brown 2021).

The following subsections describe the research process, including synthetic data generation, key feature selection, preprocessing steps, model development procedures, and evaluation metrics used to implement this research design.

Synthetic data generation

Due to the sensitive nature of the study context, the use of real patient records was not permitted. Refugee health data is governed by strict legal and ethical frameworks designed to prevent privacy breaches, discrimination, and misuse of sensitive information (ICRC 2020; UNHCR 2021). To enable predictive model development under these constraints, synthetic data was used. Methodologically, synthetic data generation is justified when access to real-world data is restricted by privacy, consent, security, or political concerns, yet analytic development and reproducibility remain necessary. Synthetic data is expected to preserve key statistical properties of the source population without exposing identifiable individuals, thereby reducing disclosure risk when compared with releasing raw microdata (El Emam, Mosquera, and Bass 2020; Rubin 1993). As Jordon, Yoon, and van der Schaar (2019) argue, synthetic data generation should be accompanied by transparent assumptions and validation procedures to demonstrate statistical fidelity and practical utility.

Following Jordon, Yoon, and van der Schaar (2019), the synthetic dataset was generated through a structured process that combined external population statistics with clinically plausible dependency rules. First, the variable schema was defined to reflect the minimum dataset required for a hospital readmission model in a humanitarian health setting. Second, marginal distributions were

parameterized from authoritative literature so that the synthetic records reflected the demographic and clinical profile of Palestinian refugees served by UNRWA through the following sources:

- Age and gender distributions were parameterized using UNRWA's official health and population statistics (UNRWA 2023).
- Chronic illness prevalence, including diabetes and hypertension, was modelled using refugee health surveillance and related healthcare literature (Kumar et al. 2025; Saria, Butte, and Sheikh 2018; WHO 2020).
- Hospital admission frequency was calibrated using evidence from humanitarian health service utilization patterns reported by Al-Jadba et al. (2024).
- Past surgeries, admission dates, and vital signs were generated from bounded clinical ranges and care patterns reported for low-resource healthcare settings (Al-Jadba et al. 2024; Makhoul, Chaiban, and Al-Hajj 2025).

Third, dependencies between variables were introduced to preserve clinically meaningful relationships, for example between chronic illness burden, prior admissions, complication risk, and readmission. This process made it possible to generate records that were realistic enough for methodological testing while remaining fully artificial.

Using the above-mentioned steps and parameters, a dataset of 1000 synthetic patient records was generated, with each record representing a unique synthetic individual rather than a real patient. In practical terms, age and gender were sampled first, chronic illnesses were then assigned using prevalence-informed probabilities, and service-use variables, admission dates, surgery histories, and vital signs were generated through rule-based and distribution-based sampling consistent with the cited literature. The target variable, readmission within 30 days, was subsequently assigned so that the dataset preserved a plausible event rate and clinically meaningful relationships between risk factors and outcomes. Microsoft Excel was used for the initial schema design and quality checks, while Python, using pandas and numpy, was used for preprocessing and model development. The key variables included in the dataset are presented in Appendix A1 and summarized below:

- *Demographics*: Age (integer) and Gender (categorical: Male/Female)
- *Medical History*: Chronic illness type (categorical), Past surgeries (numeric count)
- *Service Utilization*: Number of prior hospital admissions, Average length of stay
- *Vitals:* Blood pressure (mmHg), Glucose level (mg/dL), Heart rate (bpm)
- *Current Treatment*: Medication type (categorical), Complication risk level (categorical: Low/Medium/High)
- *Target Variable*: Readmission within 30 days (binary: 0 = No, 1 = Yes)
- *Admission Date*: Date field used for temporal analysis of hospitalizations and for possible future time-series modelling.

The above variables were selected because they capture the clinical, behavioural, and temporal signals required for ML model development. Demographic variables, such as age and gender, provide baseline stratification and support fairness checks across subgroups (James et al. 2021). Medical history captures comorbidity burden and prior clinical complexity, both of which are important predictors of future service utilization and adverse outcomes, while also improving the clinical interpretability of the model (Hastie et al. 2009). Service-utilization variables, such as prior admissions and average length of stay, reflect accumulated disease burden, continuity of care, and access patterns, all of which are often highly predictive in readmission studies (Géron 2019). Vital signs, including blood pressure, glucose, and heart rate, represent near-term physiological status and therefore improve sensitivity to acute deterioration (Bishop 2006). Current treatment and complication risk capture clinician-assessed severity and therapeutic intensity, thereby acting as useful proxies for latent clinical factors not directly observed in raw measurements (Murphy 2012). The target variable, 30-day readmission, defines the supervised learning objective and therefore requires robust evaluation because of the potential for class imbalance (James et al. 2021). Finally, admission date supports temporal validation, drift detection, and future time-aware modelling, thereby reducing leakage and improving generalisability (Hastie et al. 2009).

To examine whether the generated synthetic data satisfied the methodological expectations outlined by Bishop (2006), Géron (2019), Murphy (2012), and Jordon, Yoon, and van der Schaar (2019), the dataset was validated using multiple association tests. Pearson and Spearman correlations were computed for age, past surgeries, prior admissions, systolic blood pressure, glucose, heart rate, complication risk, admission date, and readmission. The results showed, for example, a strong Pearson association between prior hospital admissions and complication risk ($r = 0.749$). Cramér's V was also used to examine categorical associations and showed the strongest relationship between chronic illness type and medication type ($V = 0.627$). In addition, mixed associations estimated using eta indicated a strong relationship between chronic illness type and glucose level ($\eta = 0.792$). The dataset also retained a readmission prevalence of 19.5%. Details of these tests are presented in Appendices A2–A6. Collectively, these results suggest that the synthetic dataset preserved sufficient internal structure to support ML model development and evaluation.

Dataset preprocessing

In line with Géron (2019) and Murphy (2012), the dataset was preprocessed by converting each categorical variable into a machine-readable numerical representation. Binary and ordinal variables, such as gender and complication risk, were encoded directly, while other categorical variables were transformed in a way that preserved model compatibility. Numerical variables were then standardized. Z-score normalization was applied to centre and scale variables such as age, length of stay, and vital signs, ensuring that differences in feature scale did not

disproportionately influence the model. Because the dataset was synthetic, no missing values were present. Nevertheless, the codebase included conditional logic to drop or impute missing values in order to maintain robustness for future real-world applications. A final stage of feature engineering tested derived indicators, such as admission density (number of admissions divided by age), during exploratory analysis; however, these engineered variables were excluded from the final model in order to preserve interpretability. All preprocessing steps were undertaken in accordance with the recommendations of Géron (2019), Jordon, Yoon, and van der Schaar (2019), and Murphy (2012).

ML model development

Beam and Kohane (2018) and Kansagara et al. (2011) state that the predictive analytics model is expected to serve as a clinical decision-support tool, helping frontline staff in humanitarian settings identify high-risk patients for early intervention, efficient follow-up, and optimized resource use. Therefore, a modular ML pipeline was developed to predict whether a patient receiving healthcare services from UNRWA is likely to be readmitted within 30 days of discharge. The model was implemented in Python 3.12.4 and executed using plain-text scripts via the Windows command line. This environment was selected to reflect a low-infrastructure setup, allowing reproducibility in real-world humanitarian contexts where advanced computational tools may not be available as it was mentioned by Van Rossum and Drake (2009). The entire pipeline was modularized across individual scripts to improve clarity, maintainability, and operational adaptability. The pipeline was also divided into two phases, where each phase consisted of complication risk scoring, preprocessing, model training or prediction, and Excel-based outputs.

- Phase 1: Learning from historical patient data with known admission outcomes
- Phase 2: Predicting readmission risk for new patient data without known outcomes

Training of generated synthetic data

Half of the synthetic dataset was used as historical data for model training and was designed to replicate clinical features typically observed among Palestinian refugees accessing UNRWA healthcare services. Data preprocessing was carried out using the pandas and numpy libraries, as recommended by McKinney (2010), while model training and evaluation were implemented using scikit-learn, following Pedregosa et al. (2011). The model and transformation artefacts were saved for reuse using joblib. Each record in the dataset included demographic variables (age and gender), medical history (chronic illnesses, allergies, surgeries, and medications), healthcare utilization indicators (number of admissions and length of stay), vital signs (blood pressure, glucose, and heart rate), admission date, and a target variable indicating whether the patient was readmitted within 30 days, in line with Kumar et al. (2025), UNRWA (2023), WHO (2020), and Saria, Butte, and Sheikh (2018). During model training, the following steps were undertaken:

- Step 1: Assigning complication risk: A complication risk score was assigned to each patient using customized logic embedded in the script. The script is named as 'assign_complication_risk.py'. This logic considers abnormal vital signs, chronic disease presence, surgery history, and allergy load to classify each patient as Low, Medium, or High risk for post-discharge complications. These thresholds were informed by clinical guidelines and similar early warning scoring systems (Han, Kamber, and Pei 2012; Obermeyer and Emanuel 2016). The addition of this engineered feature helped the model account for underlying health severity in its predictions. A full code version is provided in Appendix A7.
- Step 2: Loading and reviewing the data: The output from Step 1, stored in the script workflow as 'load_data.py', was loaded through the pandas library following McKinney (2010). This step was used to examine completeness, data types, and column consistency, thereby ensuring that the input data was ready for transformation.
- Step 3: Preprocessing the data: In this step, the script 'data_preprocessing.py' prepared the dataset for model training. Categorical variables such as gender, chronic illness type, medication class, and complication risk level were encoded into numerical representations. For example, male was coded as 1 and female as 0, while risk levels were encoded as Low = 0, Medium = 1, and High = 2. Continuous variables such as glucose level, blood pressure, and heart rate were standardized using Z-score normalization, as advised by Kuhn and Johnson (2013), in order to avoid bias in model training arising from feature-scale disparities, as explained by Han, Kamber, and Pei (2012). The fully processed dataset was then saved as 'processed_patient_data.xlsx'. Table 1 provides the coding structure in greater detail.
- Step 4: Model training and evaluation: Model training was conducted using a script named 'model_creation.py'. The Random Forest classifier was selected because of its robustness, ability to model non-linear relationships, and relative interpretability in healthcare prediction tasks (Breiman 2001). The model was trained using 80% of the dataset, with the remaining 20% reserved for testing. The main target variable was whether the patient was readmitted within 30 days. Because class imbalance was possible, with fewer readmitted patients than non-readmitted patients, Synthetic Minority Oversampling Technique (SMOTE) was considered to reduce underrepresentation of the minority class (Chawla et al. 2002). However, SMOTE can introduce risks such as leakage, noise amplification, and unrealistic samples. To manage these risks, the workflow was designed carefully. First, data leakage was prevented by applying oversampling only to the training data, in line with Kuhn and Johnson (2013) and Lemaître, Nogueira, and Aridas (2017), who recommend that oversampling be performed within each cross-validation fold. Second, a suitable variant was selected: mixed numeric and categorical features were handled using SMOTENC, while SMOTETomek was considered for

Table 1: Mapping of categorical features to numeric values.

Feature	Category	Mapped value
Gender	Male	1
	Female	0
Chronic Illnesses	Hypertension	1
	Asthma	2
	Diabetes	3
	Heart Disease	4
	Arthritis	5
	Chronic Kidney Disease	6
	Stroke	7
	Cancer	8
	COPD	9
	Unknown	−1
Allergies	Dust	1
	Pollen	2
	Mould	3
	Animal Dander	4
	Food	5
	Shellfish	6
	Unknown	−1
Complication Risk	Low	0
	Medium	1
	High	2
	Unknown	−1

removing ambiguous points, following Chawla et al. (2002) and Lemaître, Nogueira, and Aridas (2017). Third, hyperparameters were tuned to avoid imposing an artificial 50/50 class balance and to promote calibration, as suggested by He and Garcia (2009). Finally, imbalance-robust metrics, such as precision-recall performance, F1-score, and recall, were prioritized, and probability calibration was considered when determining decision thresholds (Chicco and Jurman 2020; He and Garcia 2009; Pedregosa et al. 2011). The trained model, together with its preprocessing scaler and encoder, was saved using joblib to support future reuse in prediction tasks.

Prediction from the trained synthetic dataset

Once the model had been trained and evaluated, it was applied to the second half of the synthetic dataset. This dataset was tailored for prospective-style prediction through the following steps:

- Step 1: Complication risk scoring on new data: Using the script 'assign_complication_risk_2025.py', new patient records were evaluated using the same logic as the historical records in order to assign a complication risk level. This ensured that risk-related patterns were encoded consistently across the training and prediction datasets, as explained by Lundberg and Lee (2017).
- Step 2: Data transformation for new patients: The script 'data_preprocessing_2025.py' applied the same encoding and normalization procedures to the new dataset. Care was taken to use the original model's encoder and scaler so that the input structure remained consistent and potential distributional shifts were minimized, as recommended by Wang and Preininger (2019). The final input was saved as a structured Excel file.
- Step 3: Generating predictions: Using the script 'predict_new_data.py', the preprocessed data was passed to the trained model to estimate the likelihood of readmission within 30 days. Each patient was assigned a binary prediction value:
  - 1 = patient likely to be readmitted
  - 0 = patient not likely to be readmitted

  The final output, 'predictions_output.xlsx', included all patient features together with the prediction column, 'Predicted Readmission'. This output could be used by healthcare workers to identify and follow up patients at elevated risk, as discussed by Taddeo and Floridi (2018).

Evaluation of developed ML model

To ensure the reliability and clinical usefulness of the predictive model, the Random Forest classifier was evaluated on the held-out portion of the labelled synthetic dataset. The model was assessed using standard classification metrics, namely accuracy, precision, recall, F1-score, and ROC-AUC. These metrics are appropriate in healthcare predictive modelling because class imbalance can make performance appear stronger than it truly is when judged by accuracy alone (Chicco and Jurman 2020). Using the confusion-matrix values reported in this study (true positives = 70, true negatives = 69, false positives = 13, and false negatives = 6), the following results were obtained:

- ***Accuracy was 0.880 (88.0%)***: Accuracy measures the proportion of all predictions that were correct. In this case, 139 of 158 predictions were classified correctly. Although this indicates strong overall performance, accuracy should not be interpreted in isolation because the dataset contains fewer readmissions than non-readmissions.
- ***Precision was 0.843***: Precision measures the proportion of patients predicted to be readmitted who were actually readmitted. This means that approximately 84.3% of positive predictions were correct, indicating that the model kept false alarms at a moderate level and did not excessively over-flag patients.
- ***Recall, or sensitivity, was 0.921***: Recall measures the proportion of actual readmitted patients who were correctly identified by the model. In practical terms, the model detected about 92.1% of all readmissions, which is particularly important in clinical triage because missed high-risk patients may lead to delayed intervention.
- ***F1-score was 0.881***: The F1-score is the harmonic mean of precision and recall and is useful when the aim is to balance the cost of false positives against the cost of false negatives. ROC-AUC measures the model's ability to discriminate between positive and negative cases across decision thresholds. Based on the available classification counts, the implied hard-classification ROC-AUC was approximately 0.881, suggesting good discriminative performance. However, a fully threshold-independent ROC-AUC should ideally be calculated from predicted probabilities in future implementations. A confusion matrix was also used to visualize the classification outcomes, as presented in the following Figure 1.

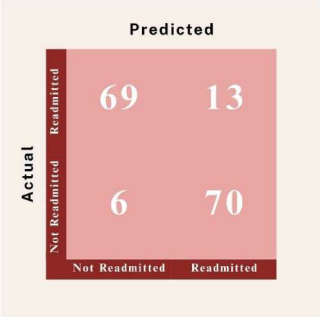

Figure 1: Confusion matrix for hospital readmission

The confusion matrix shows that 70 readmitted patients were correctly identified, and 69 patients were correctly classified as not readmitted. However, 13 patients were incorrectly flagged as readmitted (false positives), and 6 actual readmissions were missed (false negatives). This distribution indicates that the model was more effective at avoiding missed readmissions than at eliminating all false alarms, a trade-off that is often acceptable in medical triage contexts where sensitivity is prioritized (Chicco and Jurman 2020).

Taken together, the evaluation results indicate that the model performed well in predicting 30-day readmissions, particularly because of its high recall and balanced F1-score. The model therefore appears suitable as a foundation for risk-based triage and resource allocation in humanitarian healthcare settings. Nonetheless, future evaluations should also report probability-based ROC-AUC, calibration, and external validation results so that model performance can be assessed more comprehensively.

Practical implication of predictive analytics model

The metrics of predictive analytics model developed in this article show that ML can be a useful tool for forecasting hospital readmission risks in humanitarian healthcare settings. This implies that by applying predictive models to generated data reflecting the health profiles of Palestinian refugees in Syria, the developed model can be a potential tool for data-driven decision-making that can improve healthcare outcomes and resource allocation within the UNRWA health system. Similarly, it implies that the model is in line with findings from global healthcare studies conducted such as Kansagara et al. (2011). Hence, this suggests that predictive analytics models could be particularly effective in identifying high-risk patients and allowing healthcare providers to focus interventions where they are most needed. For UNRWA, this means prioritizing patients who are at greater risk of readmission, ensuring more timely follow-ups, and ultimately improving continuity of care in a context where resources are constrained. Other practical implications of the model include:

- ***Resource allocation***: Predictive models can identify regions or populations at high risk of healthcare crises, such as disease outbreaks or increased healthcare utilization, guiding the deployment of mobile health units or additional staff (Davenport and Kalakota 2019). This could also help optimize service delivery in areas with fluctuating demand.
- ***Budget prioritization***: By forecasting treatment costs for high-risk patients, predictive models could enable UNRWA to allocate funding more effectively, particularly in resource-constrained settings. For example, understanding long-term care needs for chronic disease patients could guide budget planning and help secure the necessary financial resources (Chawla et al. 2002).
- ***Public health interventions***: Predictive models can also inform public health interventions by identifying patterns of chronic disease, mental health needs, and other high-risk conditions. For example, launching chronic disease awareness campaigns in high-risk camps, where the model predicts a higher prevalence of diabetes or hypertension, could reduce future healthcare burdens (Taddeo and Floridi 2018). More examples are:
  - ***A patient flagged with high complication risk and multiple past admissions*** could be automatically referred to a follow-up team for home visits or telephonic check-ins. This could reduce unnecessary hospital readmissions and provide more personalized care.
  - ***Clinics could adjust triage protocols to prioritize high-risk patients***, reducing delays in treatment and possibly preventing a deterioration that leads to readmission. Predictive models could optimize patient flow, improving the speed and accuracy of care.
  - ***Medication adherence programmes*** could be targeted toward patients flagged by the model as having a higher likelihood of treatment failure. These patients could receive additional support, such as reminders or consultations, helping improve health outcomes and reducing hospital readmissions.
  - ***By automating aspects of patient management***, predictive models could reduce the cognitive load on healthcare providers, allowing them to focus on complex cases or emergencies.

Despite the fact that the developed model is useful, it has limitations. Firstly, while synthetic data is a powerful tool for ethical and secure model development, it is a limitation. Specifically, it may not fully capture the extreme edge cases, rare clinical conditions, or the behavioural and cultural nuances that influence real-world healthcare outcomes. In particular, individual variability in symptom reporting, treatment adherence, and external social determinants of health may be underrepresented. Secondly, some important predictors, such as environmental factors or social determinants of health (e.g. housing stability), were not included in the dataset. Incorporating such features could improve the model's accuracy and expand its applicability to real-world refugee health settings. Additionally, clinical judgment should complement the model's outputs. While predictive analytics can assist in prioritizing care, healthcare providers' expertise in contextualizing these predictions will remain crucial, especially in low-resource environments like those found in refugee populations (Taddeo and Floridi

2018). Furthermore, the model needs to address the class imbalance issue needs which can reduce the model's tendency to favour the majority class (non-readmissions). Although techniques like SMOTE were identified as potential solutions, their application in future iterations should be carefully evaluated to ensure that they do not introduce new biases or over-sample noise. In addition, while the algorithm of the model provided interpretable results, more complex models (such as deep learning methods) could be explored to better handle the intricacies of healthcare data in humanitarian settings.

Conclusion and application of predictive analytics model in real-world

So far, this article has established that the use of predictive models enables data-driven decision-making, which could lead to more effective and targeted healthcare interventions. By analyzing patterns identified through model outputs, UNRWA could make more informed decisions about the allocation of resources, budget prioritization, and public health interventions. However, it is essential to note that developing predictive model depends heavily on organizational readiness and policy alignment. The technical model itself is only one part of the solution. This is also noted by Shickel et al. (2018) that for the predictive analytics models to be effectively integrated into healthcare systems, the following components are crucial:

- ***Training programmes for clinical staff***: Healthcare workers must be equipped with the necessary knowledge and skills to interpret and act on the model's outputs. This requires targeted training to ensure that staff can use the model in their daily workflows effectively.
- ***Standard operating procedures (sops)***: Clear protocols must be established for how the model is used in practice, including data input, interpretation of results, and clinical follow-up. These SOPs will ensure consistent application and reduce errors in model usage.
- ***Feedback loops***: Predictive models must evolve as healthcare data and real-world outcomes are collected. This requires establishing feedback mechanisms to update and validate model predictions, ensuring they remain accurate over time. Feedback from frontline workers and real-time data will be critical for the refinement of these models.

Considering the studies of Fleming (2017), Bzdok et al. (2018), Saria, Butte, and Sheikh (2018) and Yozwiak, Schaffner, and Sabeti (2015) who explained how predictive analytics models could contribute to better healthcare interventions and outcomes in low-income communities, it would be conceptually important to outline how such could be applied in humanitarian settings. Thus, a framework named 'Humanitarian ML Readiness Framework (H-MLRF)' is introduced. H-MLRF is grounded in the UNRWA model case that made use of synthetic data, low-infrastructure pipelines, and strong governance needs. This framework defines 8 pillars which represent key activities that humanitarian organizations are expected to do with their deliverables. The pillars are presented in Table 2.

H-MLRF assumes that humanitarian organizations are expected to ask this readiness question: *What decision will the model support, and who acts on it?* The answer to the question enables the organization to see how the ML fits its purpose and its workflow. H-MLRF also assumes that if the ML is line with the organization's mission, the next activity is about data responsibility, privacy, and lawful basis because humanitarian health data has heightened sensitivity and political and security risk. This activity goes with data readiness and quality and Governance and risk management activities. These activities are aligned with recommendations of WHO (2021, 2019), National Institute of Standards and Technology (NIST 2023) and ISO/IEC (2023). As previously explained, humanitarian settings are low-income; thus, its technical and infrastructure readiness is constrained. Sculley et al. (2015) advises different plans such as operation, fail-safe and maintainability plan are expected to be prepared for technical fails. Similarly, Cruz Rivera et al. (2020), Liu et al. (2020), and Vasey (2022) state that internal validation and bias checks are expected to be incorporated into ML development purposely to ensure transparency and proper reporting. Moons et al. (2019) and Moons et al. (2025) also advise that using PROBAST (Prediction model Risk of Bias Assessment Tool) is important in order to assess the risk of bias and applicability of diagnostic and prognostic prediction model in ML. Training and retraining people are recommended by WHO (2021) because such efforts are essential for real-world integration.

If H-MLRF could be applied, predictive analytics models would serve as application of AI in healthcare. This corroborates the findings Zong and Guan (2025) and Gbadegeshin et al. (2025a, 2025b, 2025c) stating that ML is a key technology behind AI. Similarly, it can be concluded that AI improves efficiency and effectiveness of healthcare operations as it was mentioned by McLaughlin, Olson, and Sharma (2022), Saraswat et al. (2022); Väänänen et al. (2021). Meanwhile, it is suggested that predictive analytics models and AI should not be viewed as a replacement for human expertise but rather as a complementary decision-support tool (Alshamsi and Gbadegeshin 2025; Bhuiyan, Gbadegeshin, and Gbadegeshin 2026) because human judgment remains essential in healthcare delivery.

This article contributes to discourse on predictive analytics models, and uses of ML. It also contributes to the use of digital technologies in humanitarian healthcare systems. Thus, it is recommended that embedding predictive analytic tools within a broader digital health transformation agenda can be incorporated into humanitarian principles and once they are validated with actual refugee health records, the models could be readily adapted to UNRWA fields beyond Syria, supporting more proactive healthcare planning in Jordan, Lebanon, Gaza, and the West Bank. While the ML model is promising, it may not fully reflect the complexity of real-world clinical records. In particular, the lack of accurate historical patient data may limit the model's ability to fully learn the patterns that contribute to hospital readmissions, such

Table 2: Humanitarian ML readiness framework (H-MLRF).

Pillar	Key activities	Deliverables
1	Mission-fit and clinical workflow integration	• Decision point mapping (triage/discharge/follow-up) • Harm analysis • Human-in-the-loop Standard Operating Procedures • Escalation routes
2	Data responsibility, privacy, and lawful basis	• Data minimization • Data access controls • Incident response • Humanitarian data responsibility principles
3	Data readiness and quality	• Minimum viable dataset • Data coding standards • Data missingness strategy • Data bias checks • Data quality audits
4	Governance and risk management across the lifecycle	• Risk register (technical, clinical, social, security) • Model accountability • Approvals for each gate • Controls aligned to AI risk standards such as NIST AI RMF and ISO/IEC 23894
5	Technical and infrastructure readiness	• Offline-first operation plan • Fail-safe modes (what happens when system is down) • Computing requirements • Maintainability plan to avoid long-term 'ML technical debt'
6	Model development, validation, and transparent reporting	• Internal validation (e.g. stratified CV and temporal validation) • Subgroup performance (age/gender/site) • Calibration checks (especially if risk scores drive action) • Transparent reporting using: SPIRIT-AI (protocols), CONSORT-AI (trial reporting), and DECIDE-AI (early live clinic)
7	Clinical safety, bias, and applicability assessment	• Structured risk-of-bias • Applicability assessment using tools such as PROBAST and newer ML-focused extensions like PROBAST + AI
8	People, training, and continuous learning loops	• Role-based training (clinicians, data stewards, IT) • Model interpretation guidance • Structured feedback capture • Retraining/retirement policy

as long-term care gaps or recurrent complications. As a result, researchers and practitioners are encouraged to continue examining how different AI and predictive analytics models can be developed for humanitarian healthcare systems.

Disclosure statement

No potential conflict of interest was reported by the authors.

ORCID iDs

Mawya Mourad http://orcid.org/0009-0007-0229-9308
Olasukanmi Joseph Oladeji http://orcid.org/0009-0001-6021-0398
Gift Funmilayo Gbadegeshin http://orcid.org/0009-0004-1936-3079
Ousha Alshamsi http://orcid.org/0009-0008-1827-4807
Taiwo Fimihan Olajumoke http://orcid.org/0009-0005-1249-1593
Saheed Adebayo Gbadegeshin http://orcid.org/0000-0001-5314-6554

References

Al-Jadba, G., W. Zeidan, P. B. Spiegel, T. Shaer, S. Najjar, and A. Seita. 2024. "UNRWA at the Frontlines: Managing Health Care in Gaza during Catastrophe." *The Lancet* 403 (10428): 723–726. https://doi.org/10.1016/S0140-6736(24)00230-7.

Alshamsi, O., and S. A. Gbadegeshin. 2025, April. "AI and Project Management: Insights from Literature Review." In *Digital Horizons: Reimagining Business in the Tech Era: Proceedings of ICBT 2025*, edited by A. Bahaaeddin and A. Hamdan, Vol. 2. 297–305. Cham: Springer. https://doi.org/10.1007/978-3-032-00329-4_26.

Arowoogun, J. O., O. Babawarun, R. Chidi, A. O. Adeniyi, and C. A. Okolo. 2024. "A Comprehensive Review of Data Analytics in Healthcare Management: Leveraging Big Data for Decision-making." *World Journal of Advanced Research and Reviews* 21 (2): 1810–1821. https://doi.org/10.30574/wjarr.2024.21.2.0590.

Beam, A. L., and I. S. Kohane. 2018. "Big Data and Machine Learning in Health Care." *JAMA* 319 (11): 1317–1102. https://doi.org/10.1001/jama.2017.18391.

Bhuiyan, M. M., G. F. Gbadegeshin, and S. A. Gbadegeshin. 2026. "Uses of AI in Children's Education: A Literature Review and Future Research Agenda." In *Technovate: The Art of Business Transformation through Technology.* edited by A. Bahaaeddin and A. Hamdan, Vol. 3. 167–175. Cham: Springer. https://doi.org/10.1007/978-3-032-00264-8_18.

Bishop, C. M. 2006. *Pattern Recognition and Machine Learning*. New York: Springer.

Breiman, L. 2001. "Random Forests." *Machine Learning* 45 (1): 5–32. https://doi.org/10.1023/A:1010933404324.

Bzdok, D., Krzywinski, M. & Altman, N. 2018. "Machine Learning: Supervised Methods." *Nature Methods* 15 (1): 5–6. https://doi.org/10.1038/nmeth.4551.

Chawla, N. V., K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. 2002. "SMOTE: Synthetic Minority Over-sampling Technique." *Journal of Artificial Intelligence Research* 16:321–357. https://doi.org/10.1613/jair.953.

Chen, I. Y., Emma Pierson, Sherri Rose, Shalmali Joshi, Kadija Ferryman, and Marzyeh Ghassemi. 2021. "Ethical Machine Learning in Healthcare." *Annual Review of Biomedical Data Science* 4 (1): 123–144. https://doi.org/10.1146/annurev-biodatasci-092820-114757.

Chicco, D., and G. Jurman. 2020. "The Advantages of the Matthews Correlation Coefficient (MCC) over F1 Score and Accuracy in Binary Classification Evaluation." *BMC Genomics* 21 (1): 6. https://doi.org/10.1186/s12864-019-6413-7.

Cruz Rivera, S., et al. 2020. *SPIRIT-AI Extension* (Clinical Trial Protocol Guidance for AI Interventions).

Davenport, T., and R. Kalakota. 2019. "The Potential for Artificial Intelligence in Healthcare." *Future Healthcare Journal* 6 (2): 94–98. https://doi.org/10.7861/futurehosp.6-2-94.

El Emam, K., L. Mosquera, and J. Bass. 2020. *Practical Synthetic Data Generation: Balancing Privacy and the Broad Availability of Data*. Sebastopol: O'Reilly Media.

Fleming, K. 2017. "Predictive Analytics in Humanitarian Response: Forecasting Outbreaks during the Ebola Crisis." Humanitarian Innovation Fund. Accessed April 4, 2025. https://www.elrha.org/project/e-predictive-analytics-during-ebola/.

Fritz, A., M. Ghaffari, K. Denecke, and S. Kirrane. 2021. "Synthetic Health Data Generation." *Review and Future Directions, Journal of Biomedical Informatics* 117:103766.

Gbadegeshin, S. A. 2019a. "The Effect of Digitalization on the Commercialization Process of High-technology Companies in the Life Sciences Industry." *Technology Innovation Management Review* 9 (1): 49–63. https://doi.org/10.22215/timreview/1211.

Gbadegeshin, S. A. 2019b. "The Commercialization Process of High Technologies." PhD diss., The University of Turku.

Gbadegeshin, S. A., A. Al Natsheh, K. Ghafel, J. Tikkanen, A. Gray, A. Rimpiläinen, A. Kuoppala, J. Kalermo-Poranen, and N. Hirvonen. 2021. "What Is An Artificial Intelligence (AI): A Simple Buzzword or a Worthwhile Inevitability?" In *ICERI2021 proceedings*, 468–479. IATED.

Gbadegeshin, S. A., O. Alshamsi, T. F. Olajumoke, and G. F. Gbadegeshin. 2025a. "Artificial Intelligence in Project Risk Management: The State of Knowledge." In *Green Intelligence: Leveraging Big Data and AI for Sustainable Business Strategy*, (vol 197), edited by M. Anasweh. Cham: Springer. https://link.springer.com/book/9783032254214.

Gbadegeshin, S. A., O. Alshamsi, T. F. Olajumoke, and G. F. Gbadegeshin. 2025b. "Artificial Intelligence in Crisis Management: The Body of Knowledge from Literature." In *Green Intelligence: Leveraging Big Data and AI for Sustainable Business Strategy*, (vol 197), edited by M. Anasweh. Cham: Springer. https://link.springer.com/book/9783032254214.

Gbadegeshin, S. A., A. French, T. F. Olajumoke, I. A. A. AlZahrani, S. Y. Farah, and O. Jegede. 2025c. "Impact of Artificial Intelligence on Digital Marketing: Theoretical and Empirical Evidence." In *Tech Fusion in Business and Society. Studies in Systems, Decision and Control*, Vol. 234. edited by R. K. Hamdan, 23–33. Cham: Springer. https://doi.org/10.1007/978-3-031-84636-6_3.

Géron, A. 2019. *Hands-on Machine Learning with Scikit-learn, Keras, and TensorFlow*. 2nd ed. Sebastopol: O'Reilly Media.

Han, J., M. Kamber, and J. Pei. 2012. *Data Mining: Concepts and. Techniques*. Waltham: Morgan Kaufmann Publishers.

He, H., and E. A. Garcia. 2009. "Learning from Imbalanced Data." *IEEE Transactions on Knowledge and Data Engineering* 21 (9): 1263–1284. https://doi.org/10.1109/TKDE.2008.239.

ICRC (International Committee of the Red Cross). 2020. *Handbook on Data Protection in Humanitarian Action*. 2nd ed. Geneva: International Committee of the Red Cross. Accessed April 4, 2025. https://www.icrc.org/en/document/handbook-data-protection-humanitarian-action.

ISO/IEC. 2023. *ISO/IEC 23894:2023 — Artificial Intelligence: Guidance on Risk Management*. Accessed February 23, 2026. https://www.iso.org/standard/77304.html.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An Introduction to Statistical Learning*. 2nd ed. New York: Springer.

Jordon, J., J. Yoon, and M. van der Schaar. 2019. "PATE-GAN: Generating Synthetic Data with Differential Privacy Guarantees." *International Conference on Learning Representations (ICLR)*.

Kansagara, D., H. Englander, A. Salanitro, David Kagen, Cecelia Theobald, Michele Freeman, and Sunil Kripalani. 2011. "Risk Prediction Models for Hospital Readmission: A Systematic Review." *JAMA* 306 (15): 1688–1698. https://doi.org/10.1001/jama.2011.1515.

Kuhn, M., and K. Johnson. 2013. *Applied Predictive Modeling* (Vol. 26). New York: Springer.

Kumar, R., O. Tanous, D. Mills, B. Wispelwey, Y. Asi, W. Hammoudeh, and O. Dewachi. 2025. "Antimicrobial Resistance in a Protracted War Setting: A Review of the Literature from Palestine." *MSystems* 10 (6): e01679–24. https://doi.org/10.1128/msystems.01679-24.

Lemaître, G., F. Nogueira, and C. K. Aridas. 2017. "Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning." *Journal of Machine Learning Research* 18 (17): 1–5.

Liu, Y., P. H. C. Chen, J. Krause, and L. Peng. 2020. "Early Prediction of Sepsis Using Machine Learning Algorithms." *PLoS ONE* 15 (10): e0240835.

Lundberg, S. M., and S. I. Lee. 2017. "A Unified Approach to Interpreting Model Predictions." *NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems*, 4768–4777. https://doi.org/10.48550/arXiv.1705.07874.

Makhoul, J., L. Chaiban, and S. Al-Hajj. 2025. ""I Have Broken Bones, Where Should I Go Now?": Qualitative Research Findings on Refugees' Journey with Injury Healthcare." *Journal of Global Health Economics and Policy* 5: e2025005. https://doi.org/10.7189/001c.131756.

McKinney, W. 2010. "Data Structures for Statistical Computing in Python." *scipy* 445 (1): 51–56.

McLaughlin, D. B., J. R. Olson, and L. Sharma. 2022. *Healthcare Operations Management*. Chicago: ACHE Learn.

Moons, K. G. M., J. A. Damen, T. Kaul, L. Hooft, C. A. Navarro, P. Dhiman, A. L. Beam, et al. 2025. "PROBAST+AI: Updated Risk of Bias/Applicability Assessment for Prediction Models Using AI." *BMJ* 388:e082505.

Moons, K. G. M., Robert F. Wolff, Richard D. Riley, Penny F. Whiting, Marie Westwood, Gary S. Collins, Johannes B. Reitsma, Jos Kleijnen, and Sue Mallett. 2019. "PROBAST: A Tool to Assess Risk of Bias and Applicability of Prediction Model Studies: Explanation and Elaboration." *Annals of Internal Medicine* 170 (1): W1–W33. https://doi.org/10.7326/M18-1377.

Murphy, K. P. 2012. *Machine Learning: A Probabilistic Perspective*. Cambridge, MA: MIT Press.

NIST (National Institute of Standards and Technology). 2023. *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*. Accessed February 23, 2026. https://www.nist.gov/itl/ai-risk-management-framework.

Obermeyer, Z., and E. J. Emanuel. 2016. "Predicting the Future – Big Data, Machine Learning, and Clinical Medicine." *New England Journal of Medicine* 375 (13): 1216–1219. https://doi.org/10.1056/NEJMp1606181.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, et al. 2011. "Scikit-learn: Machine Learning in Python." *Journal of Machine Learning Research* 12:2825–2830.

Rajkomar, A., J. Dean, and I. Kohane. 2019. "Machine Learning in Medicine." *New England Journal of Medicine* 380 (14): 1347–1358. https://doi.org/10.1056/NEJMra1814259.

Rubin, D. B. 1993. "Discussion: Statistical Disclosure Limitation." *Journal of Official Statistics* 9 (2): 461–468.

Rujas, M., R. M. G. del Moral Herranz, G. Fico, and B. Merino-Barbancho. 2025. "Synthetic Data Generation in Healthcare: A Scoping Review of Reviews on Domains, Motivations, and Future Applications." *International Journal of Medical Informatics* 195:105763. https://doi.org/10.1016/j.ijmedinf.2024.105763.

Saraswat, D., P. Bhattacharya, A. Verma, V. K. Prasad, S. Tanwar, G. Sharma, Pitshou N. Bokoro, and R. Sharma. 2022. "Explainable AI for Healthcare 5.0: Opportunities and Challenges." *IEEE Access* 10:84486–84517. https://doi.org/10.1109/ACCESS.2022.3197671.

Saria, S., A. J. Butte, and A. Sheikh. 2018. "Better Medicine through Machine Learning: What's Real, and What's Artificial?" *PLOS Medicine* 15 (12): e1002721. https://doi.org/10.1371/journal.pmed.1002721.

Sculley, D., G. Holt, D. Golovin, E. Davydov, T. Phillips, D. Ebner, V. Chaudhary, M. Young, J. F. Crespo, and D. Dennison. 2015. "Hidden Technical Debt in Machine Learning Systems." *BMJ* 377:e070904.

Shaheen, M. Y. 2021. "Applications of Artificial Intelligence (AI) in Healthcare: A Review." *ScienceOpen Preprints*.

Shickel, B., P. J. Tighe, A. Bihorac, and P. Rashidi. 2018. "Deep EHR: A Survey of Recent Advances in Deep Learning Techniques for Electronic Health Record (EHR) Analysis." *IEEE Journal of Biomedical and Health Informatics* 22 (5): 1589–1604. https://doi.org/10.1109/JBHI.2017.2767063.

Smith, J., A. Jones, and T. Brown. 2021. "Socio-economic Status and Refugee Populations: A Global Overview." *Journal of Refugee Studies* 34 (3): 456–467.

Taddeo, M., and L. Floridi. 2018. "The Ethics of Artificial Intelligence: A European Perspective." *AI & Society* 33 (3): 531–538.

UNHCR (United Nations High Commissioner for Refugees). 2021. "Policy on the Protection of Personal Data of Persons of Concern to UNHCR, United Nations High Commissioner for Refugees." Accessed April 4, 2025. https://www.unhcr.org.

UNRWA (United Nations High Commissioner for Refugees). 2023. "Health Programme Annual Report, United Nations Relief and Works Agency for Palestine Refugees in the Near East." Accessed April 4, 2025. https://www.unrwa.org/resources/reports.

Väänänen, A., K. Haataja, K. Vehviläinen-Julkunen, and P. Toivanen. 2021. "AI in Healthcare: A Narrative Review." *F1000Research* 10:6. https://doi.org/10.12688/f1000research.26997.2.

Van Rossum, G., and F. L. Drake. 2009. "PYTHON 2.6 Reference Manual.".

Vasey, B., et al. 2022. "DECIDE-AI: Reporting Guideline for Early-stage Live Clinical Evaluation of AI Decision Support Systems." *BMJ* 77:e070904.

Wang, F., and A. Preininger. 2019. "AI in Health: State of the Art, Challenges, and Future Directions." *Yearbook of Medical Informatics* 28 (1): 16–26. https://doi.org/10.1055/s-0039-1677908.

WHO (World Health Organization). 2019. *Recommendations on Digital Interventions for Health System Strengthening.* Accessed February 23, 2026. https://www.who.int/publications/i/item/9789241550505.

WHO (World Health Organization). 2021. *Ethics and Governance of Artificial Intelligence for Health.* Accessed February 23, 2026. https://www.who.int/publications/i/item/9789240029200.

Wirtz, V. J., Hans V. Hogerzeil, Andrew L. Gray, Maryam Bigdeli, Cornelis P. de Joncheere, Margaret A. Ewen, Martha Gyansa-Lutterodt, et al. 2017. "Essential Medicines for Universal Health Coverage." *The Lancet* 389 (10067): 403–476. https://doi.org/10.1016/S0140-6736(16)31599-9.

World Health Organization. 2020. *Strategic Purchasing for UHC: Unlocking the Potential*. Geneva: WHO Press.

Yozwiak, N. L., S. F. Schaffner, and P. C. Sabeti. 2015. "Data Sharing: Make Outbreak Research Open Access." *Nature* 518 (7540): 477–479. https://doi.org/10.1038/518477a.

Zong, Z., and Y. Guan. 2025. "AI-driven Intelligent Data Analytics and Predictive Analysis in Industry 4.0: Transforming Knowledge, Innovation, and Efficiency." *Journal of the Knowledge Economy* 16 (1): 864–903. https://doi.org/10.1007/s13132-024-02001-z.

Appendices

Appendix A1. Distribution table for synthetic dataset

Variable	Category	Count	Percent
Age (bands)	18–29	47	4.7
Age (bands)	30–44	195	19.5
Age (bands)	45–59	335	33.5
Age (bands)	60–74	288	28.8
Age (bands)	75–95	135	13.5
Gender	Female	538	53.8
Gender	Male	462	46.2
Chronic illness type	Asthma	75	7.5
Chronic illness type	CHF	69	6.9

(*Continued*)

Continued.

Variable	Category	Count	Percent
Chronic illness type	CKD	69	6.9
Chronic illness type	COPD	86	8.6
Chronic illness type	Cancer	43	4.3
Chronic illness type	Diabetes	132	13.2
Chronic illness type	Hypertension	249	24.9
Chronic illness type	None/Other	277	27.7
Past surgeries (count)	0	330	33
Past surgeries (count)	1	345	34.5
Past surgeries (count)	2	200	20
Past surgeries (count)	3	89	8.9
Past surgeries (count)	4	28	2.8
Past surgeries (count)	5	6	0.6
Past surgeries (count)	6	2	0.2
Prior hospital admissions (bands)	0	535	53.5
Prior hospital admissions (bands)	01-Feb	291	29.1
Prior hospital admissions (bands)	03-May	160	16
Prior hospital admissions (bands)	06-Dec	14	1.4
Average length of stay (days, bands)	01-Feb	232	23.2
Average length of stay (days, bands)	Nov-20	41	4.1
Average length of stay (days, bands)	03-Apr	317	31.7
Average length of stay (days, bands)	05-Jul	315	31.5
Average length of stay (days, bands)	08-Oct	95	9.5
Systolic blood pressure (mmHg, bands)	110–129	276	27.6
Systolic blood pressure (mmHg, bands)	130–139	218	21.8
Systolic blood pressure (mmHg, bands)	140–159	348	34.8
Systolic blood pressure (mmHg, bands)	160+	100	10
Systolic blood pressure (mmHg, bands)	<110	58	5.8
Glucose (mg/dL, bands)	100–125	273	27.3
Glucose (mg/dL, bands)	126–199	237	23.7
Glucose (mg/dL, bands)	200–249	50	5
Glucose (mg/dL, bands)	250+	5	0.5
Glucose (mg/dL, bands)	<100	435	43.5
Heart rate (bpm, bands)	100–119	49	4.9
Heart rate (bpm, bands)	120+	0	0
Heart rate (bpm, bands)	60–79	487	48.7
Heart rate (bpm, bands)	80–99	398	39.8
Heart rate (bpm, bands)	<60	66	6.6
Medication type	ACE inhibitor	171	17.1
Medication type	Antibiotic	188	18.8
Medication type	Anticoagulant	55	5.5
Medication type	Beta blocker	104	10.4
Medication type	Insulin	55	5.5
Medication type	Metformin	77	7.7
Medication type	Other	91	9.1
Medication type	Statin	153	15.3
Medication type	Steroid	106	10.6
Complication risk level	High	140	14
Complication risk level	Low	600	60
Complication risk level	Medium	260	26
Readmission within 30 days	0 = No	805	80.5
Readmission within 30 days	1 = Yes	195	19.5
Admission year	2023	289	28.9
Admission year	2024	326	32.6
Admission year	2025	323	32.3
Admission year	2026	62	6.2
Admission month	April	63	6.3
Admission month	August	86	8.6
Admission month	December	91	9.1
Admission month	February	107	10.7
Admission month	January	114	11.4
Admission month	July	60	6
Admission month	June	74	7.4
Admission month	March	84	8.4
Admission month	May	73	7.3
Admission month	November	83	8.3
Admission month	October	89	8.9
Admission month	September	76	7.6

Appendix A2. Pearson correlation matrix (numeric)

	Age	Past surgeries count	Prior hospital admissions	Average length of stay days	Systolic BP_mmHg	Glucose_ mg_dL	Heart Rate_bpm	Complication RiskLevel_num	Admission Date_ordinal	Readmission Within30Days
Age	1	0.189	0.299	0.086	0.553	0.196	0.037	0.504	0.034	0.236
Past Surgeries Count	0.189	1	0.118	0.046	0.108	0.041	−0.033	0.198	0.012	0.082
PriorHospitalAdmissions	0.299	0.118	1	0.204	0.244	0.108	0.11	0.749	0.027	0.391
AverageLengthOfStayDays	0.086	0.046	0.204	1	0.053	0.034	0.032	0.453	0.03	0.219
SystolicBP_mmHg	0.553	0.108	0.244	0.053	1	0.036	0.002	0.359	0.034	0.129
Glucose_mg_dL	0.196	0.041	0.108	0.034	0.036	1	−0.062	0.312	0.036	0.099
HeartRate_bpm	0.037	−0.033	0.11	0.032	0.002	−0.062	1	0.125	0.046	0.082
ComplicationRiskLevel_num	0.504	0.198	0.749	0.453	0.359	0.312	0.125	1	0.072	0.416
AdmissionDate_ordinal	0.034	0.012	0.027	0.03	0.034	0.036	0.046	0.072	1	−0.005
ReadmissionWithin30Days	0.236	0.082	0.391	0.219	0.129	0.099	0.082	0.416	−0.005	1

Appendix A3. Top Pearson correlations (absolute)

Variance 1	Variance 2	Correlation
Prior Hospital Admissions	Complication Risk Level_num	0.749
Age	Systolic BP_mmHg	0.553
Age	Complication Risk Level_num	0.504
Average Length of Stay Days	Complication Risk Level_num	0.453
Complication Risk Level_num	Readmission Within 30Days	0.416
Prior Hospital Admissions	Readmission Within 30Days	0.391
Systolic BP_mmHg	Complication Risk Level_num	0.359
Glucose_mg_dL	Complication Risk Level_num	0.312
Age	Prior Hospital Admissions	0.299
Prior Hospital Admissions	Systolic BP_mmHg	0.244
Age	ReadmissionWithin30Days	0.236
Average Length of Stay Days	ReadmissionWithin30Days	0.219

Appendix A4. Spearman correlation matrix (numeric)

	Age	Past surgeries count	Prior hospital admissions	Average length of stay days	Systolic BP_mmHg	Glucose_mg_dL	HeartRate_ bpm	Complication RiskLevel_num	Admission Date_ordinal	Readmission Within30Days
Age	1	0.198	0.274	0.07	0.55	0.188	0.03	0.511	0.036	0.249
Past Surgeries Count	0.198	1	0.089	0.043	0.101	0.051	−0.033	0.189	0.008	0.055
Prior Hospital Admissions	0.274	0.089	1	0.208	0.198	0.104	0.104	0.691	0.041	0.351
Average Length of Stay Days	0.07	0.043	0.208	1	0.036	0.001	−0.009	0.415	0.019	0.18
Systolic BP_mmHg	0.55	0.101	0.198	0.036	1	0.052	−0.014	0.347	0.027	0.117
Glucose_mg_dL	0.188	0.051	0.104	0.001	0.052	1	−0.043	0.266	0.017	0.103
HeartRate_bpm	0.03	−0.033	0.104	−0.009	−0.014	−0.043	1	0.106	0.043	0.071
ComplicationRiskLevel_num	0.511	0.189	0.691	0.415	0.347	0.266	0.106	1	0.077	0.389
AdmissionDate_ordinal	0.036	0.008	0.041	0.019	0.027	0.017	0.043	0.077	1	−0.005
ReadmissionWithin30Days	0.249	0.055	0.351	0.18	0.117	0.103	0.071	0.389	−0.005	1

Appendix A5. Cramér's variance matrix (categorical)

Variance 1	Variance 2	Cramers variance
Chronic Illness Type	Medication Type	0.627
Chronic Illness Type	Complication Risk Level	0.456
Complication Risk Level	ReadmissionWithin30Days	0.426
MedicationType	Complication Risk Level	0.305
Chronic Illness Type	ReadmissionWithin30Days	0.295
MedicationType	ReadmissionWithin30Days	0.131
Gender	MedicationType	0.051
Gender	Chronic Illness Type	0
Gender	Complication Risk Level	0
Gender	ReadmissionWithin30Days	0

Appendix A6. Top mixed associations (Eta)

Categorical	Numeric	Eta (0-1)
Chronic Illness Type	Glucose_mg_dL	0.792
Medication Type	Glucose_mg_dL	0.788
Complication Risk Level	Prior Hospital Admissions	0.758
Complication Risk Level	Age	0.507
Chronic Illness Type	Systolic BP_mmHg	0.455
Complication Risk Level	Average Length of Stay Days	0.454
Chronic Illness Type	Prior Hospital Admissions	0.436
MedicationType	Systolic BP_mmHg	0.431
Complication Risk Level	Readmission Within 30Days	0.428
Complication Risk Level	SystolicBP_mmHg	0.359
Chronic Illness Type	Average Length of Stay Days	0.342
Chronic Illness Type	Age	0.327
Complication Risk Level	Glucose_mg_dL	0.326
Chronic Illness Type	ReadmissionWithin30Days	0.307
Chronic Illness Type	Heart Rate_bpm	0.288

Appendix A7. python codes for the full scripts

1. assign_complication_risk.py

```
import pandas as pd
# Load patient dataset
df = pd.read_excel("patient_data_2024.xlsx")
# Custom logic to flag high-risk cases based on vitals + history
def calculate_risk(row):
    if (
        row['Number_of_Admissions'] > 3 or
        row['Blood_Pressure'] > 150 or
        row['Glucose_Level'] > 200
    ):
        return 'High'
    elif (
        row['Length_of_Stay'] > 7 or
        row['Heart_Rate'] > 100
    ):
        return 'Medium'
    else:
        return 'Low'
# Add new column to the dataframe
df['Complication_Risk'] = df.apply(calculate_risk, axis = 1)
# Export excel file
df.to_excel("patient_data_2024.xlsx", index = False)
print("Complication Risk assigned and saved to 'patient_data_2024.xlsx'")
```

2. load_data.py

```
import pandas as pd
import os
# Get the directory where the script is located
script_dir = os.path.dirname(os.path.abspath(__file__))
# Build the path to the Excel file dynamically
file_path = os.path.join(script_dir, 'patient_data_2024.xlsx')
# Load the Excel file
df = pd.read_excel(file_path, sheet_name = 'Sheet1')
# Display the first few rows of the data to confirm it loaded properly
print(df.head())
```

3. data_preprocessing.py

```
import pandas as pd
import os
# Load the dataset from the patient's data file
df = pd.read_excel('patient_data_2024.xlsx')
# Check if 'Complication_Risk' column exists, raise an error if not
if 'Complication_Risk' not in df.columns:
    raise ValueError("The column 'Complication_Risk' does not exist. Please run the previous steps first
(assign_complication_risk.py and load_data.py).")
# Chronic Illnesses mapping
chronic_illnesses_mapping = {
```

```
    'Hypertension': 1,
    'Asthma': 2,
    'Diabetes': 3,
    'Heart Disease': 4,
    'Arthritis': 5,
    'Chronic Kidney Disease': 6,
    'Stroke': 7,
    'Cancer': 8,
    'COPD': 9,
    'Unknown': -1 # For unknown values
}
# Allergies mapping
allergies_mapping = {
    'Dust': 1,
    'Pollen': 2,
    'Mould': 3,
    'Animal Dander': 4,
    'Food': 5,
    'Shellfish': 6,
    'Unknown': -1 # For unknown values
}
# Complication Risk mapping
complication_risk_mapping = {
    'Low': 0,
    'Medium': 1,
    'High': 2,
    'Unknown': -1 # For unknown values
}
# Gender mapping
gender_mapping = {
    'Male': 1,
    'Female': 0
}
# Apply mappings
df['Gender'] = df['Gender'].map(gender_mapping)
df['Chronic_Illnesses'] = df['Chronic_Illnesses'].map(chronic_illnesses_mapping).fillna(-1).astype(int)
df['Allergies'] = df['Allergies'].map(allergies_mapping).fillna(-1).astype(int)
df['Complication_Risk'] = df['Complication_Risk'].map(complication_risk_mapping).fillna(-1).astype(int)
# Check the updated DataFrame
print('Updated DataFrame with Numeric Values:')
print(df.head())
# Save the cleaned data to a new Excel file called processed_patient_data.xlsx
output_file = 'processed_patient_data_2024.xlsx'
# Check current working directory
print(f"Current working directory: {os.getcwd()}")
# Save the cleaned data to a new Excel file
try:
    df.to_excel(output_file, index = False)
    print(f"\nData processing complete. The cleaned data is saved as '{output_file}'.")
except Exception as e:
    print(f"Error saving the file: {e}")
```

4. model_creation.py

```
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import classification_report, confusion_matrix
from imblearn.over_sampling import SMOTE
import joblib

# Load processed data
df = pd.read_excel('processed_patient_data_2024.xlsx')

# Drop ID column
if 'Patient_ID' in df.columns:
    df = df.drop(columns = ['Patient_ID'])

# Features and target
X = df.drop(columns = ['Readmission_Within_30_Days', 'admission_date'])
y = df['Readmission_Within_30_Days']
```

```
19
20  # Handle imbalance
21  smote = SMOTE(random_state = 42)
22  X_resampled, y_resampled = smote.fit_resample(X, y)
23
24  # Train-test split
25  X_train, X_test, y_train, y_test = train_test_split(
26      X_resampled, y_resampled, test_size = 0.2, random_state = 42
27  )
28
29  # Scale the features
30  scaler = StandardScaler()
31  X_train_scaled = scaler.fit_transform(X_train)
32  X_test_scaled = scaler.transform(X_test)
33
34  # Train the classifier
35  model = RandomForestClassifier(random_state = 42)
36  model.fit(X_train_scaled, y_train)
37
38  # Evaluate performance
39  y_pred = model.predict(X_test_scaled)
40  accuracy = model.score(X_test_scaled, y_test)
41
42  # Save model and scaler
43  joblib.dump(model, 'trained_model.pkl')
44  joblib.dump(scaler, 'scaler.pkl')
45
46  # Print evaluation metrics
47  print("Model saved as trained_model.pkl")
48  print("Accuracy:", round(accuracy, 2))
49  print("Classification Report:\n", classification_report(y_test, y_pred))
50  print("Confusion Matrix:\n", confusion_matrix(y_test, y_pred))
51
52  # Get feature importance from the trained model
53  feature_importance = model.feature_importances_
54
55  # Create a DataFrame to hold feature names and their importance
56  feature_names = X.columns
57  feature_importance_df = pd.DataFrame({
58      'Feature': feature_names,
59      'Importance': feature_importance
60  })
61
62  # Sort by importance
63  feature_importance_df = feature_importance_df.sort_values(by = 'Importance', ascending = False)
64
65  # Save to Excel
66  feature_importance_df.to_excel('feature_importance.xlsx', index = False)
67
68  # Print feature importance for review
69  print("Feature Importance:\n", feature_importance_df)
```

5. assign_complication_risk_2025.py

```
1       import pandas as pd
2
3       # Load 2025 patient data
4       df = pd.read_excel("patient_data_2025.xlsx")
5
6       # Custom logic to flag high-risk cases based on vitals + history
7       def calculate_risk(row):
8           if (
9               row['Number_of_Admissions'] > 3 or
10              row['Blood_Pressure'] > 150 or
11              row['Glucose_Level'] > 200
12          ):
13              return 'High'
14          elif (
15              row['Length_of_Stay'] > 7 or
16              row['Heart_Rate'] > 100
17          ):
18              return 'Medium'
```

```
        else:
            return 'Low'

    # Add new column
    df['Complication_Risk'] = df.apply(calculate_risk, axis = 1)

    # Save
    df.to_excel("patient_data_2025.xlsx", index = False)
    print("Complication Risk assigned and saved to 'patient_data_2025.xlsx'")
```

6. data_preprocessing_2025.py

```
    import pandas as pd
    import os

    # Load 2025 patient data
    df = pd.read_excel('patient_data_2025.xlsx')

    # Mappings
    chronic_illnesses_mapping = {
        'Hypertension': 1, 'Asthma': 2, 'Diabetes': 3, 'Heart Disease': 4,
        'Arthritis': 5, 'Chronic Kidney Disease': 6, 'Stroke': 7,
        'Cancer': 8, 'COPD': 9, 'Unknown': −1
    }
    allergies_mapping = {
        'Dust': 1, 'Pollen': 2, 'Mould': 3, 'Animal Dander': 4,
        'Food': 5, 'Shellfish': 6, 'Unknown': −1
    }
    complication_risk_mapping = {'Low': 0, 'Medium': 1, 'High': 2, 'Unknown': −1}
    gender_mapping = {'Male': 1, 'Female': 0}

    # Apply mappings
    df['Gender'] = df['Gender'].map(gender_mapping)
    df['Chronic_Illnesses'] = df['Chronic_Illnesses'].map(chronic_illnesses_mapping).fillna(−1).astype(int)
    df['Allergies'] = df['Allergies'].map(allergies_mapping).fillna(−1).astype(int)
    df['Complication_Risk'] = df['Complication_Risk'].map(complication_risk_mapping).fillna(−1).astype(int)

    # Save
    output_file = 'processed_patient_data_2025.xlsx'
    df.to_excel(output_file, index = False)
    print(f"Processed 2025 data saved as '{output_file}'")
```

7. predict_new_data.py

```
    import pandas as pd
    import joblib

    # Load the saved model and scaler
    model = joblib.load('trained_model.pkl')
    scaler = joblib.load('scaler.pkl')

    # Load the processed data
    df = pd.read_excel('processed_patient_data_2025.xlsx')

    # Save original Patient_IDs (if available)
    if 'Patient_ID' in df.columns:
        patient_ids = df['Patient_ID']
        df = df.drop(columns = ['Patient_ID'])
    else:
        patient_ids = pd.Series(range(1, len(df) + 1), name = 'Patient_ID')

    # Separate features and target
    df = df.drop(columns = [col for col in ['Readmission_Within_30_Days', 'admission_date'] if col in df.columns])


    # Scale the input features using the saved scaler
    X_scaled = scaler.transform(df)

    # Make predictions
    predictions = model.predict(X_scaled)

    # Prepare the output DataFrame
```

```
29      output_df = pd.DataFrame({
30          'Patient_ID': patient_ids,
31          'Predicted_Readmission_Within_30_Days': predictions
32      })
33
34      # Save to Excel
35      output_df.to_excel('predictions_output.xlsx', index = False)
36      print("Predictions saved to predictions_output.xlsx")
```

8. model_APP.py

```
1       import streamlit as st
2       import pandas as pd
3       import numpy as np
4       from sklearn.model_selection import train_test_split
5       from sklearn.preprocessing import StandardScaler
6       from sklearn.ensemble import RandomForestClassifier
7       from sklearn.metrics import classification_report, confusion_matrix
8       from imblearn.over_sampling import SMOTE
9       import joblib
10      import io
11      import matplotlib.pyplot as plt
12      import seaborn as sns
13
14      st.set_page_config(page_title = "UNRWA Readmission Prediction", layout = "wide")
15
16      st.title("Predictive Readmission Model for UNRWA Patients")
17
18      # File upload
19      st.sidebar.header("Upload Excel Files")
20      file1 = st.sidebar.file_uploader("Upload FIRST file (old patient data)", type = ["xlsx"])
21      file2 = st.sidebar.file_uploader("Upload SECOND file (new data to predict)", type = ["xlsx"])
22
23      start_prediction = False
24
25      if file1 and file2:
26          df1 = pd.read_excel(file1)
27          df2 = pd.read_excel(file2)
28
29          try:
30              if df1['admission_date'].min() >=  df2['admission_date'].min():
31                  st.sidebar.error("⚠ Please upload the OLDER patient data in the first upload box, and the NEWER data in the second
        box.")
32              else:
33                  st.sidebar.success("Files uploaded in correct order. Ready to proceed.")
34                  if st.sidebar.button(" Start Prediction"):
35                      start_prediction = True
36          except Exception as e:
37               st.sidebar.error(f'Error reading admission_date column. Please check your files. Details: {e}")
38
39      if start_prediction:
40          df_hist = df1
41          df_new = df2
42
43          # Assign complication risk
44          def assign_risk(df):
45              def calculate_risk(row):
46                  if (row['Number_of_Admissions'] > 3 or row['Blood_Pressure'] > 150 or row['Glucose_Level'] > 200):
47                      return 'High'
48                  elif (row['Length_of_Stay'] > 7 or row['Heart_Rate'] > 100):
49                      return 'Medium'
50                  else:
51                      return 'Low'
52              df['Complication_Risk'] = df.apply(calculate_risk, axis = 1)
53              return df
54
55          df_hist = assign_risk(df_hist)
56          df_new = assign_risk(df_new)
57
58          # Preprocessing
59          mappings = {
60              'Chronic_Illnesses': {
```

```
            'Hypertension': 1, 'Asthma': 2, 'Diabetes': 3, 'Heart Disease': 4,
            'Arthritis': 5, 'Chronic Kidney Disease': 6, 'Stroke': 7, 'Cancer': 8, 'COPD': 9
        },
        'Allergies': {
            'Dust': 1, 'Pollen': 2, 'Mould': 3, 'Animal Dander': 4, 'Food': 5, 'Shellfish': 6
        },
        'Complication_Risk': {'Low': 0, 'Medium': 1, 'High': 2},
        'Gender': {'Male': 1, 'Female': 0}
    }

    def preprocess(df, is_training = True):
        for col, mapping in mappings.items():
            df[col] = df[col].map(mapping).fillna(-1).astype(int)
        if is_training:
            return df.drop(columns = ['Patient_ID'], errors = 'ignore')
        else:
            return df.drop(columns = ['Patient_ID', 'Readmission_Within_30_Days'], errors = 'ignore')

    df_hist_processed = preprocess(df_hist, is_training = True)
    df_new_processed = preprocess(df_new, is_training = False)

    # Model training
    X = df_hist_processed.drop(columns = ['Readmission_Within_30_Days', 'admission_date'], errors = 'ignore')
    y = df_hist_processed['Readmission_Within_30_Days']

    smote = SMOTE(random_state = 42)
    X_resampled, y_resampled = smote.fit_resample(X, y)

    X_train, X_test, y_train, y_test = train_test_split(X_resampled, y_resampled, test_size = 0.2, random_state = 42)

    scaler = StandardScaler()
    X_train_scaled = scaler.fit_transform(X_train)
    X_test_scaled = scaler.transform(X_test)

    model = RandomForestClassifier(random_state = 42)
    model.fit(X_train_scaled, y_train)
    y_pred = model.predict(X_test_scaled)

    # Prediction on new data
    new_scaled = scaler.transform(df_new_processed.drop(columns = ['admission_date'], errors = 'ignore'))
    new_predictions = model.predict(new_scaled)
    df_new['Predicted_Readmission'] = new_predictions

    # Display performance
    st.subheader("Model Evaluation Metrics")
    accuracy = model.score(X_test_scaled, y_test)
    report = classification_report(y_test, y_pred, output_dict = True)
    matrix = confusion_matrix(y_test, y_pred)

    st.write(f"**Accuracy:** {accuracy:.2f}")
    st.write("**Classification Report:**")
    st.dataframe(pd.DataFrame(report).transpose())

    st.write("**Confusion Matrix:**")
    fig, ax = plt.subplots()
    sns.heatmap(matrix, annot = True, fmt = 'd', cmap = 'Blues', xticklabels = ['No', 'Yes'], yticklabels = ['No', 'Yes'])
    ax.set_xlabel("Predicted")
    ax.set_ylabel("Actual")
    st.pyplot(fig)

    # Feature Importance
    st.subheader("Feature Importance")
    importances = model.feature_importances_
    feature_names = X.columns
    imp_df = pd.DataFrame({'Feature': feature_names, 'Importance': importances}).sort_values(by = 'Importance', ascending
= False)
    st.bar_chart(imp_df.set_index('Feature'))

    # Show feature importance as a table
```

```
129     st.write("### Feature Importance Table")
130     st.dataframe(imp_df.style.format({"Importance": "{:.4f}"}))
131
132     # Prediction Results
133     st.subheader("Prediction Results on New Data")
134     st.dataframe(df_new[['admission_date', 'Age', 'Gender', 'Predicted_Readmission']].head(10))
135
136     # Save to local disk (optional)
137     df_new.to_excel("predicted_results_2025.xlsx", index = False)
138
139     # Download prediction results
140     output = io.BytesIO()
141     with pd.ExcelWriter(output, engine = 'xlsxwriter') as writer:
142         df_new.to_excel(writer, index = False, sheet_name = 'Predictions')
143     st.download_button("Download Full Prediction Results", data = output.getvalue(), file_name
    = "predicted_results_2025.xlsx", mime = "application/vnd.openxmlformats-officedocument.spreadsheetml.sheet")
144 else:
145     st.info("Please upload both the historical and new patient Excel files to continue.")
```

9. shap_explainer.py

```
1   import shap
2   import matplotlib.pyplot as plt
3   import pandas as pd
4   import joblib
5
6   # Load trained model and scaler
7   model = joblib.load("trained_model.pkl")
8   scaler = joblib.load("scaler.pkl")
9
10  # Load the 2024 processed dataset
11  df = pd.read_excel("processed_patient_data_2024.xlsx")
12
13  # Select feature columns (drop target, ID, and date)
14  X = df.drop(columns = ["Readmission_Within_30_Days", "Patient_ID", "admission_date"], errors = 'ignore')
15
16  # Scale the features using the same scaler used in training
17  X_scaled = scaler.transform(X)
18  X_scaled_df = pd.DataFrame(X_scaled, columns = X.columns)
19
20  # Create SHAP explainer
21  explainer = shap.TreeExplainer(model)
22
23  # Compute SHAP values for scaled features
24  shap_values = explainer.shap_values(X_scaled_df)
25
26  # Safely select correct SHAP values shape
27  if isinstance(shap_values, list) and len(shap_values) == 2 and shap_values[1].shape[0] == X_scaled_df.shape[0]:
28      shap_vals = shap_values[1]
29  else:
30      shap_vals = shap_values
31
32  # Plot summary bar chart
33  plt.clf()
34  plt.close('all')
35  # Use SHAP values for class 1 (readmitted)
36  shap.summary_plot(shap_vals[:, :, 1], X_scaled_df, plot_type = "bar", show = False)
37
38  print("SHAP value shape:", shap_vals.shape)
39  print("Feature matrix shape:", X_scaled_df.shape)
40  print("Feature names:", list(X_scaled_df.columns))
41
42  # Save chart
43  plt.savefig("shap_summary_bar_chart.png", dpi = 300, bbox_inches = 'tight')
44  print("SHAP summary bar chart saved as 'shap_summary_bar_chart.png'")
```