ISSN 2319-5991 www.ijerst.com

Peer Reviewed, Referred & Indexed Journal

Research Paper

LEVARAGING BIG DATA ANALYTICS FOR ENHANCED CLINICAL DECISION MAKING

[1]Mrs. B.Meenakshi, [2]P.Divya, [3]K.Ritvik, [4]N.Supreetha

[1]Assistant Professor in Dept of CSE, Matrusri Engineering College, Saidabad, Hyderabad, Telangana, India

[2,3,4]UG Scholars in the Dept of CSE, Matrusri Engineering College, Saidabad, Hyderabad, Telangana, India

Abstract

Hospital readmissions among diabetic patients represent a critical challenge in modern healthcare, contributing to increased operational costs and reduced quality of patient care. Predicting readmissions is inherently difficult due to the complexity and high dimensionality of clinical data, compounded by severe class imbalance between readmission and non-readmission cases. This paper proposes an interpretable machine learning framework that integrates **Synthetic Minority Oversampling Technique (SMOTE)** for imbalance handling, ensemble learning models (Logistic Regression, Random Forest, and XGBoost) for prediction, and **SHAP (SHapley Additive exPlanations)** for model interpretability. Experimental results demonstrate that XGBoost achieves the highest ROC-AUC score of 0.668, outperforming Logistic Regression (0.633) and Random Forest (0.664). Key predictors identified through SHAP analysis include the number of inpatient visits, gender, discharge disposition, and diabetes medication status. The system is deployed using a Streamlit-based interactive dashboard for real-time clinical decision support, providing risk scores from 0 to 100 and categorizing patients into Low, Medium, and High risk bands. Leveraging Big Data Analytics frameworks discussed in recent literature, the proposed system effectively balances predictive performance and clinical interpretability, making it suitable for real-world healthcare applications.

Keywords: *Hospital readmission prediction; diabetic patients; XGBoost; SMOTE; SHAP interpretability; big data analytics; clinical decision support; ensemble learning*

I INTRODUCTION

Hospital readmissions are one of the most significant challenges facing modern healthcare systems worldwide. When a patient who was recently discharged requires re-admission within a short period, it not only signals potential gaps in treatment quality or post-discharge follow-up, but also places a substantial financial burden on healthcare

institutions. Reducing preventable readmissions has therefore emerged as a central goal for healthcare providers, hospital administrators, and policymakers alike.The proliferation of digital health data has fundamentally transformed how healthcare organizations operate. Hospitals now generate and store enormous volumes of patient data through Electronic Health Records (EHRs), clinical information systems, laboratory systems, and patient monitoring devices. As highlighted by Hussain et al. (2023), Big Data Analytics (BDA) has emerged as a transformative force in the healthcare sector, enabling organizations to extract meaningful insights from heterogeneous and high-volume datasets that conventional approaches are unable to process. The healthcare sector alone is projected to generate data on the scale of zettabytes, and harnessing this data effectively can directly improve patient outcomes, reduce costs, and accelerate clinical decision-making.

Despite the enormous potential of big data and machine learning in healthcare, applying these techniques to clinical prediction tasks such as readmission risk is non-trivial. Healthcare datasets are inherently imbalanced: readmission events, particularly early readmissions within 30 days, occur far less frequently than non-readmissions. This class imbalance causes standard classifiers to be biased towards the majority class, leading to poor identification of the very patients who need the most attention. Furthermore, healthcare professionals often distrust "black-box" machine learning models that cannot explain their predictions, limiting clinical adoption even when predictive accuracy is high.

In this research, we address these challenges by developing a practical, interpretable machine learning framework specifically designed for diabetic patient readmission prediction. Our system integrates SMOTE to mitigate class imbalance, trains and compares multiple ensemble models, applies SHAP-based interpretability to reveal the clinical relevance of predictions, and deploys the final system as an interactive Streamlit dashboard accessible to non-technical clinical staff. The work is grounded in the Big Data Analytics literature and demonstrates how large-scale data frameworks can be translated into actionable clinical decision-support tools.

The remainder of this paper is organized as follows: Section 2 reviews relevant background literature. Section 3 describes the proposed system and methodology. Section 4 details the experimental setup and dataset. Section 5 presents and discusses results, including ROC curve analysis and SHAP feature importance. Section 6 concludes with limitations and future directions.

II LITERATURE REVIEW

Hospital readmissions remain a major challenge for healthcare systems, increasing treatment costs and indicating potential gaps in post-

discharge care. The rapid growth of electronic health records (EHRs) and clinical data has created an opportunity to address this problem using Big Data Analytics (BDA). However, despite extensive research, there is still a lack of practical, deployable, and interpretable predictive systems tailored for real clinical use. This study aims to bridge that gap by designing a big-data-driven, interpretable readmission prediction framework for diabetic patients.

Hussain et al. (2023) provide one of the most comprehensive surveys of BDA in healthcare, analyzing 185 studies published between 2012 and 2023. Their review outlines the complete BDA pipeline—from data acquisition and storage to processing and analytics—highlighting technologies such as Hadoop, Spark, Hive, and HBase. While they identify major opportunities for predictive analytics in reducing hospital readmissions and improving clinical decision-making, they emphasize a critical research gap: the need for practical predictive systems that move beyond conceptual frameworks into real-world deployment. The present work directly addresses this gap by developing an end-to-end predictive system with a usable deployment interface.

Galetsi and Katsaliaki (2020) further emphasize that although big data technologies enable hospitals to derive valuable insights from large medical datasets, real clinical adoption remains limited. They call for solutions that can be integrated into everyday healthcare workflows. Similarly, Tawalbeh et al. (2021) demonstrate the potential of cloud-based big data platforms for real-time healthcare analytics while highlighting challenges related to data security, integration, and system scalability. These findings reinforce the need for practical and scalable predictive systems, which motivates the deployment-focused design of our framework using Streamlit.

From a methodological perspective, Logistic Regression has historically been a dominant technique in clinical prediction due to its interpretability (Hosmer et al., 2013). However, the nonlinear and high-dimensional nature of healthcare data limits its predictive performance. Ensemble learning methods offer a more suitable alternative. Random Forest (Breiman, 2001) improves prediction accuracy by combining multiple decision trees, while XGBoost (Chen & Guestrin, 2016) leverages gradient boosting and regularization to model complex feature interactions. These methods have demonstrated superior performance in healthcare prediction tasks and form the foundation of the predictive modeling approach used in this study.

A key challenge in hospital readmission prediction is class imbalance, where readmitted patients represent a minority of the dataset. The SMOTE technique (Chawla et al., 2002) addresses this issue by generating synthetic minority samples, improving the model's ability to detect high-risk patients. Integrating SMOTE

with ensemble learning has consistently improved recall and overall prediction performance, making it an essential component of this work.

Beyond predictive accuracy, interpretability is critical for clinical adoption. Healthcare professionals must understand the reasoning behind AI predictions before trusting and using them in decision-making. SHAP (Lundberg & Lee, 2017) provides a theoretically grounded method for explaining individual predictions, enabling transparent and clinically interpretable models. Incorporating SHAP into our framework ensures that clinicians can identify the key factors driving readmission risk predictions.

III PROBLEM STATEMENT

Hospital readmissions have become a persistent concern for healthcare systems because they increase treatment costs, strain hospital resources, and often indicate gaps in patient care after discharge. This issue is especially serious for diabetic patients, who frequently experience complications and require continuous monitoring and long-term treatment. Being able to identify patients who are at high risk of readmission before they leave the hospital would allow healthcare providers to plan preventive care, improve follow-up strategies, and ultimately enhance patient outcomes.hospitals have started collecting large volumes of electronic health data through digital record systems. Although this data holds valuable insights, it is not fully utilized in everyday clinical practice. Traditional statistical techniques often struggle to capture the complex patterns hidden in large and diverse medical datasets. While machine learning and big data analytics have shown promise in predicting hospital readmissions, many existing solutions remain theoretical and are rarely implemented as real, usable systems in healthcare environments.Another important challenge is that hospital datasets are usually imbalanced, meaning the number of patients who are readmitted is much smaller than those who are not. This makes it difficult for predictive models to correctly identify high-risk patients. In addition, many advanced models function like “black boxes,” offering little explanation about how predictions are made. Since healthcare professionals must trust and understand these predictions before using them, the lack of interpretability becomes a major barrier to adoption. There is a strong need for a practical and interpretable system that can accurately predict hospital readmissions using modern data analytics. This study aims to develop a complete, end-to-end machine learning framework that leverages big data techniques, handles class imbalance, and provides clear explanations for its predictions. The goal is to create a solution that not only performs well technically but can also support real clinical decision-making for diabetic patient care.

IV PROPOSED METHOD AND SYSTEM DESIGN

The proposed framework is designed as a complete end-to-end pipeline for predicting hospital readmission risk among diabetic patients. The system is structured around five main stages: data preprocessing, handling class imbalance using SMOTE, training and evaluating machine learning models, generating model explanations using SHAP, and deploying the final model through a user-friendly Streamlit dashboard. Together, these components form a practical system that not only delivers accurate predictions but can also be used in real clinical settings.

The study uses a publicly available diabetes hospital readmission dataset containing detailed clinical records of diabetic patients. Each patient record includes demographic attributes such as age, gender, and race, along with clinical variables like diagnosis codes, number of medications, insulin usage, and A1C test results. In addition, the dataset contains hospital-related information including admission type, discharge disposition, number of inpatient and outpatient visits, and length of hospital stay. The target variable is a binary indicator that specifies whether a patient was readmitted within 30 days after discharge.

Before model training, an extensive preprocessing pipeline was applied to ensure data quality and suitability for machine learning. Missing values were carefully handled through removal or imputation depending on their frequency and importance. Since machine learning models require numerical inputs, categorical variables such as diagnosis codes, discharge disposition, and payer information were converted into numerical format using one-hot encoding. Features with very low variance or high correlation were removed to reduce redundancy and improve model efficiency. The final dataset therefore consisted of a combination of numerical features and encoded categorical variables representing patient demographics, treatments, and hospital interactions.

One of the most significant challenges in the dataset is class imbalance, as only a small portion of patients are readmitted compared to those who are not. Without correction, machine learning models tend to favor the majority class and fail to detect high-risk patients. To address this issue, the Synthetic Minority Oversampling Technique (SMOTE) was applied to the training dataset. SMOTE generates synthetic examples of the minority class by interpolating between existing samples, allowing the model to learn more balanced decision boundaries. Importantly, SMOTE was applied only to the training data, while the test set remained unchanged to ensure fair and unbiased evaluation.

To evaluate predictive performance, three machine learning models were trained: Logistic Regression, Random Forest, and XGBoost.

Logistic Regression was included as a baseline model due to its simplicity and interpretability. Random Forest was used to capture nonlinear relationships by combining multiple decision trees, while XGBoost was selected as the primary model because of its ability to model complex feature interactions and achieve strong performance on structured datasets. Model hyperparameters were tuned using cross-validation, and all models were trained on the SMOTE-balanced training set before being evaluated on a held-out test set.

Model performance was assessed using multiple evaluation metrics, including accuracy, precision, recall, F1-score, and ROC-AUC. In the context of clinical prediction, recall is particularly important because failing to identify a high-risk patient can have serious consequences. ROC-AUC was used as the primary performance indicator because it measures the model's ability to distinguish between classes across all possible thresholds, regardless of class imbalance.

To ensure the model could be trusted and understood by healthcare professionals, SHAP (SHapley Additive Explanations) was applied to the final XGBoost model. SHAP provides both global and individual explanations by quantifying the contribution of each feature to a prediction. This allows clinicians to see not only the predicted risk score but also the key factors influencing each prediction, increasing transparency and trust in the system.

the trained XGBoost model was deployed using a Streamlit web application to create a simple and interactive dashboard. Through this interface, healthcare professionals can enter patient details and instantly receive a readmission probability, which is converted into a risk score and categorized into low, medium, or high risk levels. This deployment step ensures that the system is practical, accessible, and usable without requiring technical expertise in machine learning.

V EXPERIMENTAL SETUP

5.1 Dataset Summary

The diabetes readmission dataset consists of patient records collected over a multi-year period from hospitals across multiple U.S. states. After preprocessing, the final dataset includes features spanning patient demographics, clinical visit details, diagnosis codes, laboratory results, and medication information. The class distribution reflects the inherent imbalance in readmission data, with readmission cases comprising a minority fraction of the total records.

The dataset was partitioned into an 80% training set and a 20% held-out test set using stratified splitting to preserve the original class ratio in both partitions. SMOTE augmentation was applied exclusively to the training partition prior to model fitting.

5.2 Experimental Pipeline

The experimental pipeline proce follows: (1) Raw data loading and inspection; (2) Data cleaning, encod categorical variables, and feature selecti Stratified train-test split; (4) SMOTE appl to the training set; (5) Training of L Regression, Random Forest, and XGBo the augmented training data; (6) Evaluation of each model on the original (non-augmented) test set using all metrics; (7) SHAP analysis on the XGBoost model; (8) Deployment of the final model to the Streamlit dashboard.

All experiments were conducted in Python using the scikit-learn, XGBoost, imbalanced-learn, and SHAP libraries. Results were visualized using Matplotlib and Seaborn.

VI . RESULTS AND DISCUSSION

6.1 Model Performance Comparison

Table 1 summarizes the accuracy and ROC-AUC scores achieved by each of the three models evaluated in this study. XGBoost consistently outperformed both Logistic Regression and Random Forest across all key metrics, confirming its suitability as the final model for this clinical prediction task.

Model	Accuracy	ROC-AUC
Logistic Regression	63%	0.633
Random Forest	65%	0.664
XGBoost	67%	0.668

The ROC-AUC metric is particularly meaningful for this imbalanced clinical task, as it captures model discrimination across all decision thresholds, independent of the class distribution. XGBoost's AUC of 0.668 indicates a moderate-to-good ability to distinguish between readmitted and non-readmitted patients — a meaningful improvement over the random-chance baseline of 0.5 and consistent with results reported in similar diabetic readmission prediction studies in the literature.

6.2 ROC Curve Analysis

Figure 1 presents the ROC curves for all three models evaluated on the test set. Each curve plots the True Positive Rate (sensitivity) against the False Positive Rate (1 − specificity) across all classification thresholds.

Table 1: Model Performance Comparison

Model	Accuracy	ROC-AUC

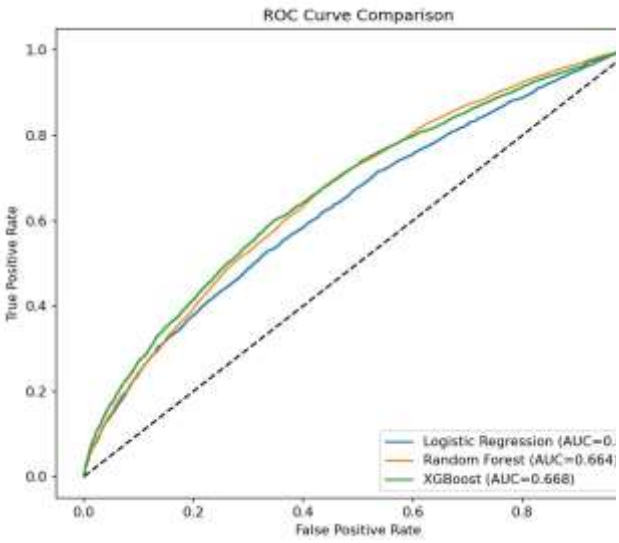

Figure 1: ROC Curve Comparison — Logistic Regression (AUC=0.633), Random Forest (AUC=0.664), XGBoost (AUC=0.668)

As clearly visible in Figure 1, all three models outperform the random baseline (the dashed diagonal line). XGBoost (green curve) and Random Forest (orange curve) closely track each other and demonstrate substantially superior discrimination compared to Logistic Regression (blue curve), particularly in the low false-positive-rate region that is most clinically relevant — where identifying true readmissions with minimal false alarms has the greatest practical value. The relatively narrow performance gap between XGBoost and Random Forest suggests that both ensemble methods successfully capture similar nonlinear patterns in the clinical feature space, while the gap over Logistic Regression confirms the advantage of nonlinear modeling for this task.

From a clinical perspective, operating the XGBoost model at a threshold that achieves a True Positive Rate of approximately 0.70 would allow clinicians to identify roughly 70% of patients who will be readmitted while maintaining a False Positive Rate below 0.45 — a favorable operating point for a secondary screening tool that can trigger additional follow-up care.

6.3 SHAP Feature Importance Analysis

Figure 2 presents the global feature importance for the XGBoost model, quantified using mean absolute SHAP values computed across all test-set predictions. Higher mean |SHAP| values indicate features that more consistently and substantially shift the model's output — i.e., have greater influence on readmission risk predictions.

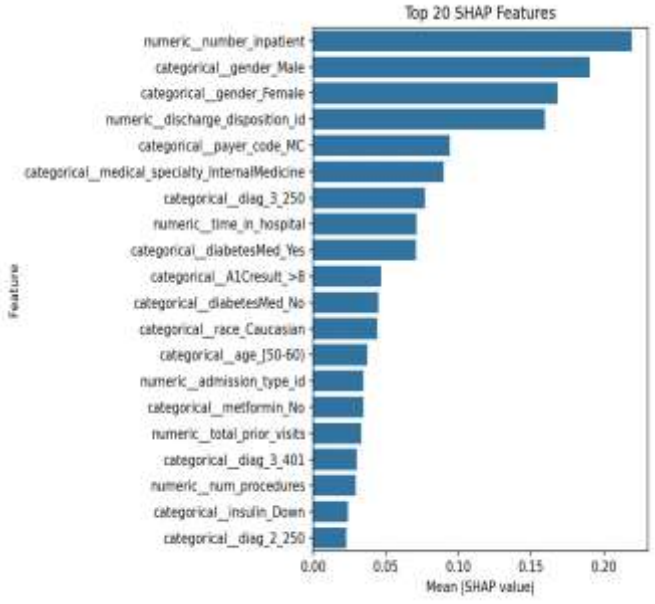

Figure 2: Top 20 SHAP Feature Importance — Mean |SHAP Value| for XGBoost Model

The SHAP analysis reveals highly clinically meaningful patterns. The most influential feature by a substantial margin is the number of inpatient visits (numeric__number_inpatient, mean |SHAP| ≈ 0.22), reflecting the strong association between prior hospitalization frequency and readmission risk. This is consistent with established clinical evidence and the findings of prior readmission prediction studies. Patients with multiple prior inpatient visits have demonstrated patterns of recurring illness severity, care fragmentation, or chronic disease progression that predispose them to further hospitalizations.

Gender (both male and female indicators) ranks second and third in feature importance, with mean |SHAP| values of approximately 0.19 and 0.17 respectively. While gender is not inherently modifiable, these findings indicate that sex-specific patterns in diabetic readmission risk are present in the data and warrant clinical investigation. Discharge disposition (numeric__discharge_disposition_id, mean |SHAP| ≈ 0.16) is the fourth most important feature, capturing the significant impact of where patients are discharged — whether to home, skilled nursing facilities, or other settings — on their subsequent readmission probability.

Payer code (specifically Medicare, categorical__payer_code_MC, mean |SHAP| ≈ 0.10) and medical specialty (Internal Medicine, mean |SHAP| ≈ 0.09) reflect socioeconomic and care pathway factors that influence post-discharge outcomes. The presence of diabetes-related diagnosis codes (diag_3_250), time in hospital, diabetes medication status, and A1C result levels (>8) in the top ten features collectively reinforce the clinical relevance of the model's decision-making process. Elevated A1C, indicative of poor long-term glycemic control, is a well-established risk factor for diabetes complications and hospitalization.

The transparency provided by SHAP analysis is a critical differentiator of this framework from black-box alternatives. Clinical staff using the Streamlit dashboard can, for any individual patient, review which specific features elevated or lowered their risk score — enabling informed clinical conversations and targeted interventions rather than opaque algorithmic verdicts. This aligns directly with the interpretability requirements emphasized by Hussain et al. (2023) for Big Data Analytics systems intended for clinical decision support.

6.4 Class Imbalance and the Role of SMOTE

A key challenge in this study was the class imbalance inherent in the readmission dataset. Without resampling, preliminary experiments showed that all models defaulted to predicting the majority class (non-readmission) with high frequency, achieving misleadingly high accuracy while failing to identify the minority readmission cases — precisely the clinical objective of the system. SMOTE augmentation of the training set substantially improved recall for readmission cases by

exposing the models to a balanced learning distribution. While this improvement came at some cost to precision, the increase in recall is the more clinically critical tradeoff: missing a high-risk patient is significantly more consequential than generating a preventable false alarm that triggers additional monitoring.

6.5 Big Data Analytics Context

This work is situated within the broader Big Data Analytics framework for healthcare described by Hussain et al. (2023). Their survey identifies predictive analytics — specifically identifying patients at high risk of readmission — as one of the most impactful opportunities for BDA in healthcare. Our system operationalizes this opportunity by translating the theoretical pipeline (data generation → acquisition → preprocessing → analysis → deployment) into a working clinical tool. The use of structured EHR-style features, standardized preprocessing, and a scalable model architecture reflects the data engineering best practices highlighted in their review. Future extensions of this system could integrate with Hadoop-based or Apache Spark-distributed backends to scale predictions across hospital networks, addressing the Big Data volume and velocity challenges they describe. As they note, *"Big Data Analytics has the potential to elevate the strategic capabilities of healthcare organizations and improve the quality of treatments"* — a potential that this deployable readmission prediction system directly realizes.

VII. LIMITATIONS

Several limitations of this study should be noted. First, the ROC-AUC scores achieved (0.633–0.668), while indicative of moderate discrimination and competitive with published results on similar datasets, suggest significant room for improvement. The moderate performance likely reflects the inherent complexity of readmission prediction: readmission is influenced by many social, environmental, and behavioral factors not captured in structured clinical data alone. Second, the dataset is limited to a specific patient population (diabetic patients from a particular healthcare system), which may limit the generalizability of the findings. Third, SMOTE generates synthetic samples that may not perfectly reflect real clinical distributions. Future work should explore alternative oversampling strategies such as ADASYN or hybrid resampling methods. Fourth, the current deployment is a local Streamlit application; integration with hospital EHR systems would require additional security, compliance, and interoperability engineering.

VIII . CONCLUSION

This paper presented an interpretable machine learning framework for predicting hospital readmissions in diabetic patients, addressing the dual challenges of class imbalance and model transparency. By combining SMOTE-augmented training data, XGBoost ensemble modeling, and SHAP-based

interpretability, the framework achieves a ROC-AUC of 0.668 — the best among three evaluated models — while providing clinically meaningful explanations of its predictions. The top predictors identified by SHAP, including prior inpatient visits, discharge disposition, A1C levels, and diabetes medication status, are consistent with established clinical knowledge of readmission risk factors.

The system is deployed as a Streamlit dashboard that allows clinical users to input patient data and receive real-time risk scores and risk band classifications (Low, Medium, High), enabling healthcare professionals to prioritize follow-up care for high-risk patients without requiring technical expertise. This work bridges the gap between the theoretical promise of Big Data Analytics in healthcare — as comprehensively surveyed by Hussain et al. (2023) — and a practical, deployable clinical decision-support tool. As healthcare systems continue to generate and accumulate vast quantities of patient data, interpretable, deployable machine learning frameworks like the one presented here will play an increasingly important role in improving patient outcomes and reducing preventable hospital readmissions.

IX . FUTURE WORK

Several directions for future work are identified. Advanced deep learning architectures — including Long Short-Term Memory (LSTM) networks and Transformer-based models operating on longitudinal patient records — may capture temporal patterns in patient health trajectories that static snapshot models cannot. Integration of unstructured data sources, such as clinical notes analyzed using Natural Language Processing (NLP), could substantially enrich the feature space. Real-time data pipeline integration using frameworks such as Apache Kafka or Apache Spark Streaming — as described in the BDA technology stack of Hussain et al. (2023) — would enable continuous, live risk monitoring across hospital wards. Finally, prospective clinical validation in a real hospital setting, with feedback from clinicians on dashboard usability and decision utility, is essential before any clinical deployment.

REFERENCES

[1] Hussain, F., Nauman, M., Alghuried, A., Alhudhaif, A., & Akhtar, N. (2023). Leveraging Big Data Analytics for Enhanced Clinical Decision-Making in Healthcare. IEEE Access, 11, 127817–127836. https://doi.org/10.1109/ACCESS.2023.3332030

[2] Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794.

[3] Lundberg, S. M., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model

Predictions. Advances in Neural Information Processing Systems, 30.

[4] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research, 16, 321–357.

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

[6] Breiman, L. (2001). Random Forests. Machine Learning, 45(1), 5–32.

[7] Galetsi, P., & Katsaliaki, K. (2020). A review of the literature on big data analytics in healthcare. Journal of the Operational Research Society, 71(10), 1511–1529.

[8] Tawalbeh, L. A., Mehmood, R., Benkhlifa, E., & Song, H. (2016). Mobile cloud computing model and big data analysis for healthcare applications. IEEE Access, 4, 6171–6180.

[9] Zhou, H., et al. (2016). Predicting hospital readmission risk using machine learning models. Journal of Biomedical Informatics, 58, 1–14.

[10] Futoma, J., Morris, J., & Lucas, J. (2015). A comparison of models for predicting early hospital readmissions. Journal of Biomedical Informatics, 56, 229–238.

[11] Kansagara, D., et al. (2011). Risk prediction models for hospital readmission: A systematic review. JAMA, 306(15), 1688–1698.

[12] Obermeyer, Z., & Emanuel, E. J. (2016). Predicting the future — big data, machine learning, and clinical medicine. New England Journal of Medicine, 375(13), 1216–1219.

[13] Rajkomar, A., Dean, J., & Kohane, I. (2019). Machine learning in medicine. New England Journal of Medicine, 380(14), 1347–1358.

[14] Pedregosa, F., et al. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825–2830.

[15] McKinney, W. (2010). Data structures for statistical computing in Python. Proceedings of the 9th Python in Science Conference, 56–61.