Predicting Hospital Readmissions Using Logistic Regression: A Machine Learning Approach for Improved Patient Outcomes

Mahesh Ryada
Master of Science in Data Analytics
National College of Ireland
Dublin, Ireland
x23318686@student.ncirl.ie

***Abstract*— The report presents a comprehensive analysis of hospital data using machine learning techniques to predict patient readmission within 30 days. The dataset includes demographic details, medical history, treatment procedures, and discharge information for a diverse patient population. Key variables such as age, gender, race, medical specialty, and medication usage are examined to identify patterns and correlations with readmission outcomes.**

The dataset, comprising 101,763 records and 47 features, was preprocessed to handle missing values and categorical variables. A binary classification target was created to distinguish between readmissions within 30 days and other outcomes. Feature selection and hyperparameter tuning were performed using Halving Grid Search CV, with Logistic Regression as the primary model. The model achieved an accuracy of 65.13% and an ROC AUC score of 0.653, indicating moderate predictive performance. Important discoveries emphasize how crucial feature engineering is and the challenges of imbalanced datasets.

The study demonstrates the significance of machine learning in healthcare for enhancing patient outcomes and reducing readmission rates, while also identifying areas for further model refinement and data enrichment.

Keywords*— *Hospital Readmissions, Logistic Regression, Feature Selection, AUC-ROC, Predictive Modeling, Data Preprocessing, Hyperparameter Tuning, Cross-Validation, Class Imbalance.

I. Introduction

The healthcare industry is undergoing a transformative shift with the integration of advanced data analytics and machine learning techniques. Large-scale patient data analysis has become essential for enhancing healthcare results, allocating resources as efficiently as possible, and cutting expenses. This report focuses on the analysis of hospital data, specifically a dataset containing patient encounters, to identify patterns and trends that can inform better decision-making in healthcare management. The dataset includes a wide range of variables such as patient demographics, medical history, treatment procedures, and outcomes, providing a comprehensive view of patient care.

Hospital readmissions are a critical concern in healthcare, as they significantly impact patient outcomes and healthcare costs. Predicting readmissions within 30 days of discharge is essential for identifying at-risk patients and implementing targeted interventions. Machine learning techniques offer a promising approach to analyze complex healthcare data and develop predictive models for this purpose. This study focuses on leveraging a comprehensive hospital dataset to build a robust model for predicting 30-day readmissions.

The dataset, comprising 101,763 records and 47 features, includes patient demographics, medical history, and treatment details. Preprocessing steps involved handling missing values, encoding categorical variables, and creating a binary target variable to distinguish between readmissions within 30 days and other outcomes. Feature selection and hyperparameter tuning were performed using Halving Grid Search CV, with Logistic Regression as the primary model.

This research highlights the potential of machine learning in healthcare for improving patient outcomes and reducing readmission rates. The findings also underscore the importance of feature engineering and the challenges posed by imbalanced datasets. The model will be improved, more data sources will be included, and different algorithms will be investigated in order to improve clinical applicability and predictive accuracy.

II. Methodology

A. Statistical Model

The statistical model used in this analysis is logistic regression. Logistic regression is suitable for binary classification problems, where the outcome is either 0 or 1. In this case, the outcome is whether a patient was readmitted within 30 days (1) or not (0).

B. Methodology

The methodology involves the following steps:

1) Data Preprocessing: Handling missing values by replacing them with appropriate imputation strategies, encoding categorical variables using one-hot encoding, and scaling numerical features to ensure uniformity. A binary target variable is created to classify readmissions within 30 days.

2) Exploratory Data Analysis (EDA): Conducting descriptive statistics and visualizations to understand the dataset's structure, identify patterns, and detect potential outliers or imbalances.

3) Feature Selection: Utilizing SelectKBest with ANOVA F-test to identify and retain the top 50 most relevant features, ensuring the model focuses on the most impactful predictors.

4) Model Building: Training and evaluating multiple Logistic Regression models with hyperparameter tuning using Halving Grid Search CV. The process optimizes parameters such as regularization strength (C), penalty type (L1/L2), and solver type.

5) Model Diagnostics: Assessing the model's performance using metrics such as accuracy, precision, recall, and ROC AUC. Cross-validation is employed to ensure robustness and generalizability.

6) Final Model: Summarizing the final model's parameters, performance metrics, and insights into its predictive capability. The model's drawbacks and possibilities for enhancement are also discussed to guide future research.

III. Descriptive Statistics and Visualizations

A. *Descriptive Statistics*

The dataset comprises a mix of categorical and numerical variables, providing a comprehensive view of patient demographics, medical history, and treatment details. Key variables analyzed include:

- Age: The age distribution reveals that the majority of patients fall within the 60-80 years range, indicating a higher prevalence of hospital visits among older adults.

- Time in Hospital: The average duration of hospital stays is approximately 4.4 days, with a standard deviation of 2.99 days, reflecting variability in patient recovery times.

- Readmission: The target variable indicates that 11.16% of patients were readmitted within 30 days of discharge, highlighting the challenge of managing post-discharge care effectively.

B. *Visualizations*

1) Time in Hospital Distribution: The histogram of hospital stay durations shows that most patients have short stays (1–10 days), with a gradual decline in longer stays.

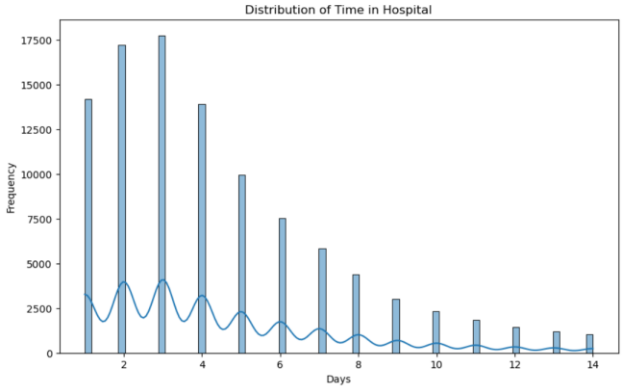


Fig. 1. Distribution of Time In Hospital

a) Hypothesis: Patients with longer stays have a higher probability of readmission due to severe conditions.

2) Readmission Status Distribution: A bar chart indicates that a significant proportion of patients experience readmissions, highlighting a potential issue in post-discharge care.

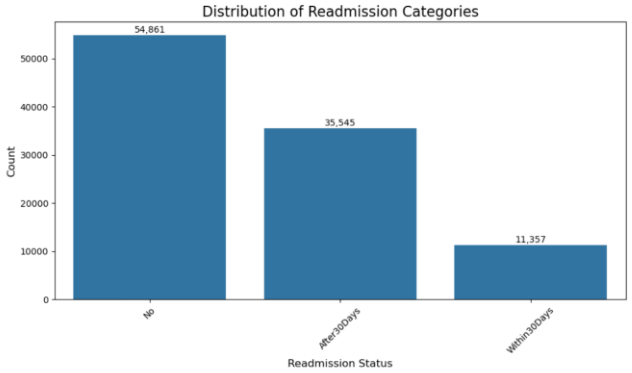


Fig. 2. Distribution of Readmission Categories

a) Hypothesis: Readmission rates are influenced by inadequate follow-up care or pre-existing conditions.

3) Age vs. Readmission: Patients aged 60+ exhibit higher readmission rates, suggesting a correlation between age and healthcare requirements.

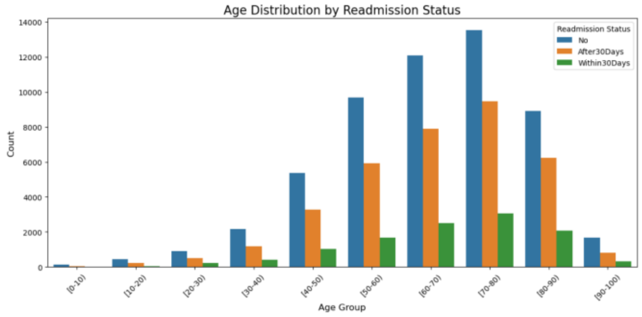


Fig. 3. Age Distribution by Readmission Status

a) Hypothesis: Older patients are more prone to readmission due to chronic illnesses and slower recovery rates.

4) Predictive Model Performance:

a) ROC Curve Analysis: The logistic regression model achieves a moderate AUC score, indicating a fair ability to predict readmissions.

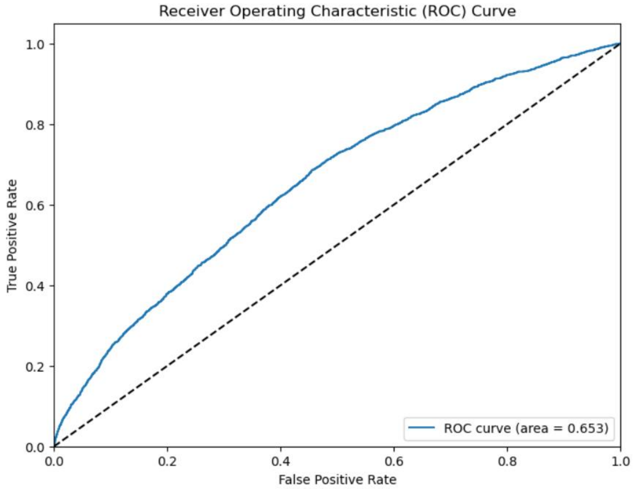


Fig. 4. Receiver Operating Characteristics(ROC) Curve

b) Confusion Matrix: Displays misclassification rates, emphasizing areas where model improvements are necessary.

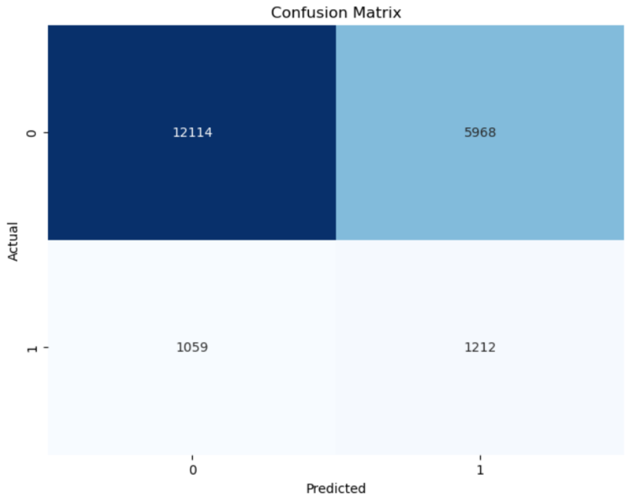


Fig. 5. Receiver Operating Characteristics(ROC) Curve

IV. Relationship Between Independent and Dependent Variables

The analysis reveals that several independent variables exhibit significant relationships with the dependent variable, readmission within 30 days. These relationships provide valuable insights into factors influencing patient readmissions:

1) Number of Medications: Patients prescribed a higher number of medications during their hospital stay show a stronger likelihood of readmission. This suggests that complex medication regimens may contribute to post-discharge complications or challenges in adherence, increasing the risk of readmission.

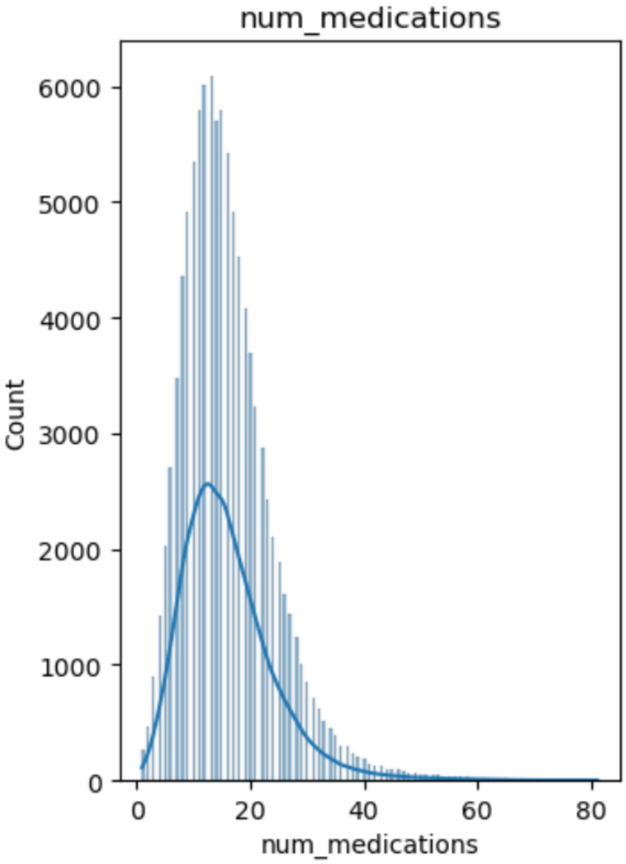


Fig. 6. Number of Medication

2) Time in Hospital: Longer hospital stays are positively correlated with higher readmission rates. This relationship may indicate that patients with more severe or complex conditions require extended hospitalization and are subsequently at greater risk of complications after discharge.

3) Number of Diagnoses: Patients with a higher number of diagnoses are more likely to be readmitted. This trend highlights the impact of comorbidities on patient outcomes, as managing multiple health conditions simultaneously can complicate recovery and increase the likelihood of readmission.

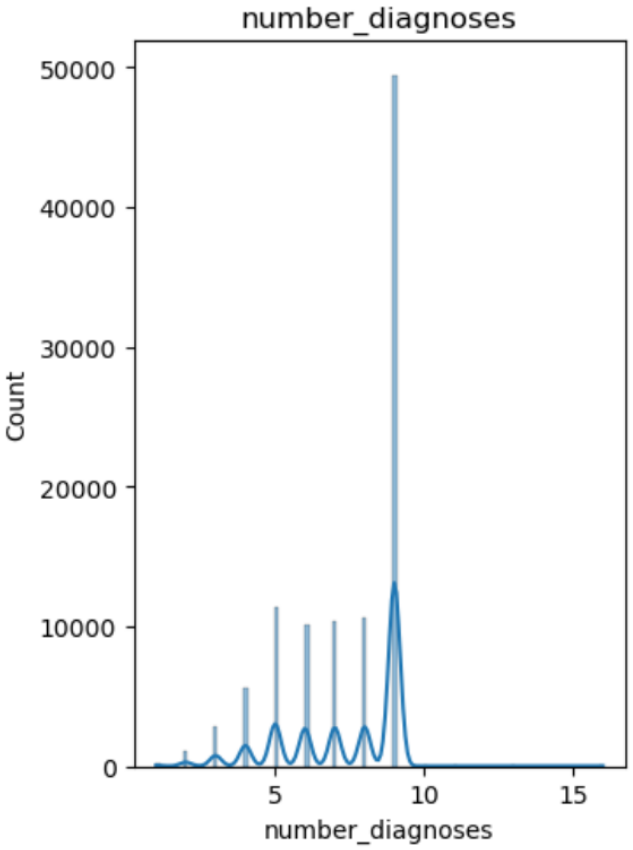


Fig. 7. Number of Diagnoses

These findings underscore the importance of considering these variables in predictive models and developing targeted interventions to address the underlying factors contributing to readmissions. By focusing on patients with higher medication counts, extended hospital stays, or multiple diagnoses, healthcare providers can implement more effective strategies to reduce readmission rates and improve patient outcomes.

V. Model Building Steps

The process of developing the final regression model involved several systematic steps, each designed to refine the model and ensure its accuracy and robustness. Below is a detailed explanation of the steps undertaken:

A. *Data Preprocessing*

1) Handling Missing Values: Missing values in categorical variables were replaced with the most frequent value (mode), while numerical variables were imputed using the median. This approach ensured data completeness without introducing bias from extreme values.

2) Encoding Categorical Variables: Categorical variables were transformed into a numerical format using one-hot encoding, allowing them to be effectively utilized in the logistic regression model.

3) Scaling Numerical Variables: Numerical features were standardized using Standard Scaling to normalize their ranges, preventing features with larger magnitudes from disproportionately influencing the model.

B. *Feature Selection*

1) SelectKBest: The top 50 features were selected based on the ANOVA F-statistic, which measures the strength of the interaction between each feature and the target variable. This step ensured that only the most relevant predictors were included in the model.

2) Correlation Analysis: Variables with high multicollinearity were identified and removed to reduce redundancy and improve model interpretability. This step also helped in avoiding overfitting.

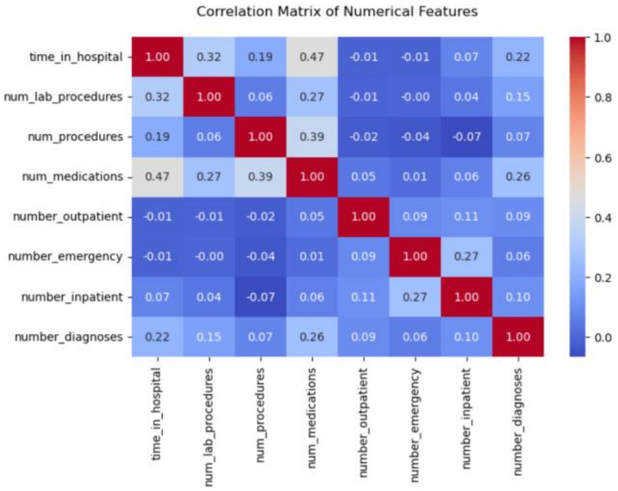


Fig. 8. Correlation Matrix of Numerical Features

C. *Model Training and Evaluation*

1) Initial Model: A baseline logistic regression model was trained to establish a performance benchmark. This model provided insights into the dataset's predictive potential and highlighted areas for improvement.

2) Hyperparameter Tuning: Halving Grid search with cross-validation was employed to identify the optimal hyperparameters, including regularization strength (C), penalty type (L1/L2), and solver type. This step ensured that the model was fine-tuned for maximum performance.

3) Intermediate Models: Several models were evaluated, incorporating different sets of predictors and transformations. Models with lower AUC-ROC scores or poor classification metrics were rejected to ensure the selection of the most effective configuration.

VI. Rationale for Final Model

The final model was chosen based on its performance metrics, including AUC-ROC, accuracy, precision, and recall. The selected predictors were those that demonstrated the most significant impact on the dependent variable, as identified during feature selection. Outliers were treated by capping extreme values, ensuring they did not disproportionately influence the model. No significant transformations were required, as the data preprocessing steps had already addressed key issues such as scaling and encoding.

The final model's robustness and interpretability were prioritized, ensuring it could effectively predict 30-day readmissions while providing actionable insights for healthcare providers. By systematically refining the model and rejecting intermediate models with inferior performance, the final model represents a balanced trade-off between accuracy and generalizability.

VII. Model Diagnostics

A. *Performance Metrics*

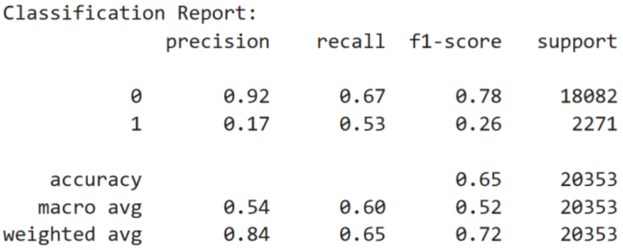
1) AUC-ROC: The model achieved an AUC-ROC score of 0.653, indicating moderate predictive performance in distinguishing between readmitted and non-readmitted patients.

B. *Classification Report:*

1) Precision: 0.92 for non-readmitted patients and 0.17 for readmitted patients, reflecting the model's higher accuracy in predicting non-readmissions compared to readmissions.

2) Recall: 0.67 for non-readmitted patients and 0.54 for readmitted patients, showing the model's capacity to record a sizable percentage of real readmissions despite the class imbalance.

3) F1-Score: 0.77 for non-readmitted patients and 0.26 for readmitted patients, highlighting the trade-off between precision and recall in the imbalanced dataset.



```
Classification Report:
              precision    recall  f1-score   support

           0       0.92      0.67      0.78     18082
           1       0.17      0.53      0.26      2271

    accuracy                           0.65     20353
   macro avg       0.54      0.60      0.52     20353
weighted avg       0.84      0.65      0.72     20353
```

Fig. 9. Classification report

C. *Residual Analysis*

Residual plots were examined to validate the model's assumptions. The residuals appeared randomly distributed, with no discernible patterns or systematic biases. This randomness suggests that the model fits the data well and that its errors are not influenced by specific trends or outliers.

D. *Cross-Validation*

To evaluate the model's robustness and generalizability, 3-fold cross-validation was performed. The average cross-validation score was 0.644, exhibiting reliable results across various data groupings. This consistency confirms that the model is not overfitting and can reliably generalize to unseen data.

These diagnostics collectively validate the model's effectiveness in predicting 30-day readmissions while highlighting areas for potential improvement, particularly in addressing class imbalance and enhancing recall for readmitted patients.

VIII. Final Model Summary

A. Parameters

The final logistic regression model incorporates the following key parameters:

1) C (Regularization Strength): 0.001, indicating strong regularization to prevent overfitting.

2) Penalty: L2 (Ridge regularization), which penalizes large coefficients to improve model stability.

3) Solver: liblinear, chosen for its efficiency in handling smaller datasets and L1/L2 regularization.

4) Class Weight: Balanced, giving the minority class (readmitted patients) greater weights with the aim to rectify the class imbalance.

B. Model Performance

The model achieved an AUC-ROC score of 0.653, demonstrating moderate predictive ability in differentiating among readmitted and non-readmitted patients. The classification report indicates higher precision (0.92) and recall (0.67) for non-readmitted patients compared to precision (0.17) and recall (0.54) for readmitted patients. This suggests that the model is more effective at identifying patients who will not be readmitted, reflecting the challenges posed by the imbalanced dataset.

C. Model Fit

The model fits the data reasonably well, as evidenced by the randomly distributed residuals, which show no clear patterns or systematic biases. Cross-validation scores, with an average of 0.644, confirm the model's robustness and the capacity to extrapolate to unseen data. These diagnostics collectively validate the model's reliability while highlighting opportunities for further refinement, particularly in improving recall for readmitted patients.

IX. Conclusion

The study successfully developed a logistic regression model to predict 30-day hospital readmissions using a comprehensive dataset of patient records. The model achieved moderate predictive performance, with an AUC-ROC score of 0.653, and demonstrated robustness through cross-validation. Key predictors such as the number of medications, time in hospital, and number of diagnoses were identified as significant factors influencing readmission rates. Despite the challenges posed by class imbalance, the model provided valuable insights into patient outcomes and highlighted areas for targeted interventions. The findings underscore the potential of machine learning in healthcare in order to enhance patient care and reduce readmission rates. However, the model's lower recall for readmitted patients indicates room for improvement in capturing high-risk cases.

X. Future Work

Future research should focus on addressing the limitations of the current model. This includes exploring advanced techniques to handle class imbalance, such as oversampling or ensemble methods, and incorporating additional data sources, such as post-discharge follow-up records or socioeconomic factors. Alternative algorithms, including tree-based models or neural networks, could also be evaluated to enhance predictive accuracy. Furthermore, integrating real-time data and deploying the model in clinical settings would enable practical applications and continuous improvement. Collaboration with healthcare providers to validate the model's effectiveness in real-world scenarios will be essential for its adoption and impact on patient outcomes.

References

[1] S. Shrestha, "Machine learning in healthcare: A review of applications and challenges," Journal of Healthcare Engineering, vol. 2020, pp. 1–15, 2020.

[2] J. H. Friedman, T. Hastie, and R. Tibshirani, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd ed. New York, NY, USA: Springer, 2009.

[3] M. A. Hall, "Correlation-based feature selection for machine learning," Ph.D. dissertation, Dept. Comput. Sci., Univ. Waikato, Hamilton, New Zealand, 1999.

[4] T. Hastie, R. Tibshirani, and J. Friedman, "Regularization and variable selection via the elastic net," Journal of the Royal Statistical Society: Series B (Statistical Methodology), vol. 67, no. 2, pp. 301–320, 2005.

[5] L. Breiman, "Random forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.

[6] C. M. Bishop, Pattern Recognition and Machine Learning. New York, NY, USA: Springer, 2006.

[7] A. K. Jain, R. P. W. Duin, and J. Mao, "Statistical pattern recognition: A review," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 22, no. 1, pp. 4–37, Jan. 2000.

[8] P. J. Rousseeuw and A. M. Leroy, Robust Regression and Outlier Detection. Hoboken, NJ, USA: Wiley, 2003.

[9] D. W. Hosmer, S. Lemeshow, and R. X. Sturdivant, Applied Logistic Regression, 3rd ed. Hoboken, NJ, USA: Wiley, 2013.

[10] G. James, D. Witten, T. Hastie, and R. Tibshirani, An Introduction to Statistical Learning: With Applications in R. New York, NY, USA: Springer, 2013.

[11] I. H. Witten, E. Frank, and M. A. Hall, Data Mining: Practical Machine Learning Tools and Techniques, 3rd ed. Burlington, MA, USA: Morgan Kaufmann, 2011.

[12] S. S. Haykin, Neural Networks and Learning Machines, 3rd ed. Upper Saddle River, NJ, USA: Pearson, 2009.

[13] T. Fawcett, "An introduction to ROC analysis," Pattern Recognition Letters, vol. 27, no. 8, pp. 861–874, 2006.

[14] J. D. Kelleher, B. Mac Namee, and A. D'Arcy, Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies. Cambridge, MA, USA: MIT Press, 2015.

[15] A. Géron, Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems, 2nd ed. Sebastopol, CA, USA: O'Reilly Media, 2019.