Interdisciplinary Journal of Computing & AI

IJCAI
Interdisciplinary Journal of Computing & AI

IJCAI

RESEARCH

Predictive Analysis of Readmission Risk of Heart Failure Patients within 28 days

Sahith Aitha[1], Muhammad Adib Uz Zaman[2*]

[1]*School of Information Technology, University of Cincinnati, Ohio, United States of America; aithash@mail.uc.edu*
[2]*School of Information Technology, University of Cincinnati, Ohio, United States of America; adib.zaman@uc.edu*

ARTICLE INFO.

Keywords:
Class Imbalance, Electronic Health Records, Ensemble Learning, Heart failure, Machine Learning, Readmission Prediction

*Corresponding author email address: adib.zaman@uc.edu

Manuscript received: 10 December 2025/
Accepted: 12 January 2026
Published online: 13 January 2026
https://doi.org/10.5281/zenodo.18228018

Cite as: S. Aitha and M. Adib Uz Zaman, 'Predictive Analysis of Readmission Risk of Heart Failure Patients within 28 days', Interdisciplinary Journal of Computing & AI (IJCAI), vol. 1, no. 1. Aspiration Publishing Trust (DARPAN ID: WB/2025/0887371), p. 1-12, Jan. 2026. doi: 10.5281/zenodo.18228018.

ABSTRACT

Heart failure (HF) is one of the emerging global health issues and a leading cause of readmission as well as a source of healthcare system strain due to escalating expenses. Proper identification of patients with high risks of readmission early, especially within 28 days, presents an opportunity to intervene before the patient is readmitted. In our paper, we suggest a machine learning-based model to forecast the risk of early readmission with the help of structured electronic health record (EHR) data. We put different preprocessing methods, SHAP-based feature pruning, and cost-sensitive learning into our pipeline, as well as tuning thresholds. We tested and trained several models like Support Vector Machines (SVM), XGBoost, LightGBM, and a Stacked Ensemble model. The stacked model, with threshold tuning and SHAP interpretability, appeared superior in terms of the cases of high-risk detection. Our findings indicate that the use of interpretability and recall-optimized thresholding would provide a fair and clinically feasible readmission prediction approach.

1 Introduction

Heart failure is one of the common reasons for hospital stays, costing the US healthcare system over $30 billion each year. Each one out of five patients (126,059, or 20.2%) returned to the hospital within a month of leaving, typically around 12 days after their initial discharge from the hospital [1]. A significant burden is being posed on healthcare systems due to the hospital's readmissions within 28 days. Predicting readmission risk can promote efficient patient management, and cost reduction. There is a huge advancement of technology, especially in Artificial Intelligence and Machine learning. Applying these to process Clinical data in the form of Electronic Medical Records (EMR), would create a significant impact in health industry and enable pro-active patient care [2].

However, various issues pertaining to imbalance in the minority class, low interpretability of models and generalizability have made it difficult to deploy many predictive models. This paper suggests a data-based predictive model that combines state-of-the-art ML models, data preprocessing strategies, threshold optimization, and model explanation by SHAP (SHapley Additive exPlanations). Our proposed methodology emphasizes the usefulness of a stacked optimal strategy in attaining improved recall rates of the minority class in the population, which is a group of patients that are likely to be readmitted. Thus, this makes the framework suitable for practical clinical applications.

2 Related Work

The issue of early hospital readmission prediction of heart failure (HF) patients has been a popular one that has been studied within the statistical, machine learning and deep learning framework. Logistic regression and Cox proportional hazards have been employed because they are simple and interpretable. These models however tend not to effectively capture the non-linear interactions of complex EHR data, and do not perform well when exposed to imbalanced datasets. Over the past few years, ensemble approaches such as Random Forest, XGBoost, and LightGBM have proven to be more predictively effective. Chen et al. [3] developed a predictive model for 6-month heart failure readmissions using data from approximately 2000 patients. Zhang et al. [1] have used XGBoost and SHAP (SHapley Additive exPlanations) values to predict the 30-day readmission of acute heart failure patients; they have an AUC greater than 0.76 and feature transparency is one of the benefits of their model.

Wu and colleagues designed a machine learning-based prognostic outcome model that was utilized to forecast mortality of people with comorbidities (heart failure and atrial fibrillation; HF- AF) in three years based on data concerning

558 patients [4]. Following the process of feature selection with Boruta and LASSO, 6 ML models were trained and tested with many available metrics. CatBoost achieved the best result where its AUC was 0.809 and F1-score was 0.715 on the test set. According to SHAP analysis, there were several predictors, including NYHA, ALC, hs-CRP, BNP, and age. The study showed that interpretable ML, especially when using SHAP explanations and other ensemble models, such as CatBoost, can consider stratifying a long-term risk and personalize treatment in patients with HF-AF.

A large body of research on the topic of ML applied to HF readmission has been thoroughly reviewed by Yu and Son [5]. Their results show that ensemble learners tend to perform better than linear models, but without interpretability and insufficient management of class imbalance, they still remain unlikely to be adopted in clinical practice. Moreover, most of the studies are optimized based on accuracy or AUC, and this is usually at the cost of recall, which is an important parameter when it comes to healthcare applications. Cost-sensitive learning and threshold tuning have been discussed in several papers as methods of optimizing the detection of minority classes.

In a similar manner, it was recently shown by Zahar [6] that SHAP-based feature pruning generated more understandable outputs and improved the performance of the hospital outcome prediction model. Although this has been done, a gap in this area is left in terms of combining threshold tuning, feature pruning, ensemble learning, interpretability, and making this into a single, clinically relevant pipeline. This paper fills this gap by suggesting a systematic ML framework that aims to maximize recall and transparency based on real-world EHR data.

A comparative analysis of SMOTE, cost-sensitive learning (CSL) and threshold tuning in predicting cardiovascular morbidity using the dataset MIMIC-III has been carried out by Zahar et al. [7]. Their paper showed that CSL would be slightly better than the others almost all the time, with the target tuning technique showing modest gains and a deterioration of the performance when using SMOTE (and mostly in ANN-based models). The results obtained by the authors are consistent with those we received as cost-sensitive modification and threshold optimization became ascendant in terms of increasing minority class recall and model robustness compared to synthetic sampling.

3 Methods

3.1 Dataset Description

The data analyzed in the current study was obtained at one of the hospitals in Sichuan, China, and consists of electronic health records (EHRs) of 2,008 heart-failure patients that were admitted to the hospital between 2016 and

2019 [8]. The data is arranged in 3 files: dat.csv (demographics, vitals, diagnostics), dat_md.csv (medication history), and metadata file which contains definitions of column names. These files were combined and tidied to create an ultimate organized table that is ready to be used as an ML training file. The target variable is a binary variable, which indicates the presence/absence of readmission within 28 days after discharge (Figure 1).

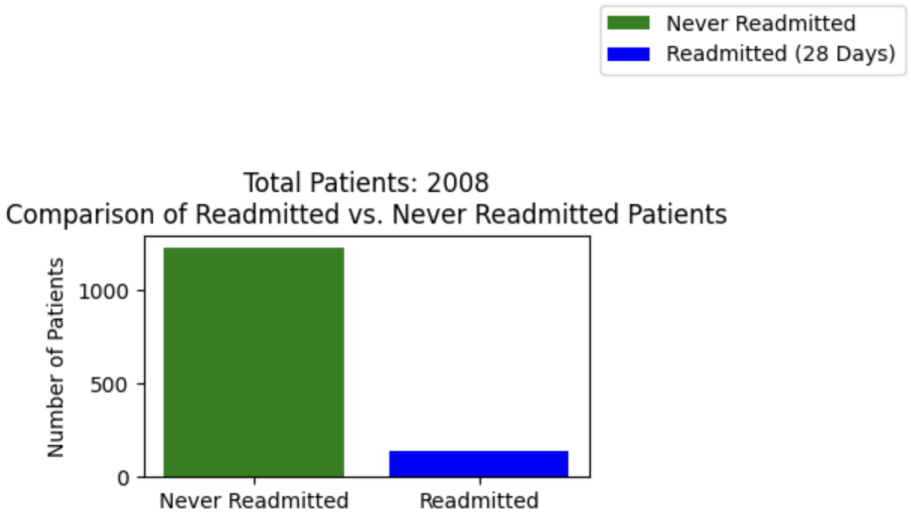

Fig. 1 Dataset Overview

All the data about patients contained in this study had been thoroughly de-identified beforehand. The data had been collected according to the existing data privacy laws, and no personal data with reference to the identifiable data were utilized throughout the study.

3.2 Data Preprocessing

Various preprocessing tasks were executed, to guarantee data quality and ML-readiness:

- **Missing Value Handling:** Missing values have been filled in using an Iterative Imputer as a continuous variable since it was effective in maintaining feature variance. The most common strategy was used to impute categorical variables.
- **Encoding:** Ordinal fields like NYHA class, Killip grade, level of consciousness were encoded with a label, and nominal fields such as gender, type of heart failure were one-hot encoded.
- **Medication Embedding**: These medication lists were tokenized and padded to fixed lengths to have a chance of application in the future in deep learning pipelines.

- **Standardization:** All the numeric features were standardized through z-score normalization.
- **Feature Truncation:** The fields like occupation, admission ward, and mortality indicators were truncated either because these fields were clinically irrelevant or could lead to data leakage as per your domain annotations.

3.3 Feature Selection and Dimensionality Reduction

To decrease dimensionality, we first used Recursive Feature Elimination (RFE). Subsequently, SHAP values were calculated over several folds in to sort features according to their significance. The features with lower than mean SHAP importance were dropped, leading to over 120 non-important variables being discarded. To solve the high dimensionality of a large-scale experiment further, we tried Principal Component Analysis (PCA) with 95 percent retention of variance. Although PCA was not included in the final model because of the loss in interpretability, it allowed optimizing the run time and memory consumption in exploratory runs.

3.4 Handling Class Imbalance

Given the heavy class imbalance (approximately 80% of the cases are not readmitted, 20% of them are readmitted), multiple strategies were evaluated:

- SMOTE, ADASYN, and SMOTE-Tomek were tested, but often led to marginal or unstable gains, especially for tree-based models. A study [9] compared five resampling methods (e.g., SMOTE, Borderline-SMOTE, NearMiss) thoroughly and the best strategies on those methods of threshold tuning on credit risk. The experiments they have done with eight classifiers showed that threshold tuning especially the tuning of G-Mean can work as well as or better than more conventional resampling strategies, and there was little added gain when resampling was also used with it. This agrees with what we found, as threshold tuning in isolation achieved considerable gains in recall and F1-score on highly imbalanced clinical datasets, which offers evidence of its practical value in deployment scenario with little Z-score standardization (ZSS) augmentation [9].
- Cost-sensitive learning was more impactful, particularly with SVM and XGBoost. For instance, adjusting class_weight to 'balanced' in SVM improved recall from 6% to over 70%. Enembreck, Loezer and Barddal, complemented by Britto, handled the same issue by proposing Cost-sensitive Adaptive Random Forest (CSARF), which assigns varying misclassification costs to the various classes in the data stream situation

[10]. Their experiment proved that cost-based strategy can be highly effective in the imbalanced classification task even without massive data resampling to acquire high recall and F1-score. This supports effectiveness of the use of class weights and cost-sensitive loss functions such as applied in our model, especially in enhancing minority class detection in a healthcare setting such as the one we have examined, the heart failure readmission.

- Threshold tuning emerged as the most effective and model-agnostic technique. By adjusting the decision to threshold from 0.5 to 0.4, we significantly improved the recall of the minority class without introducing synthetic samples.

3.5 Machine Learning Models

To thoroughly test the readmission prediction, we constructed and optimized five different machine learning model configurations, each with their own advantages in terms of processing structured clinical data and class imbalance (Figure 2):

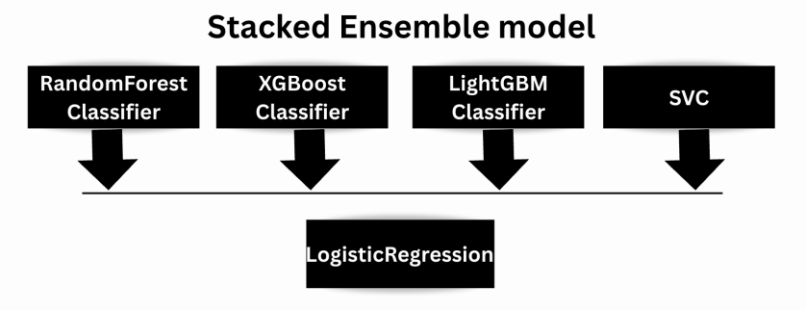

Fig. 2 Stacked Ensemble Classifier

- **Random Forest Classifier:** We also added to our stacked ensemble framework a Random Forest classifier as a base learner. To fix the class's imbalance training issue, the model was set to use a "balanced" class weight. To be more specific, we had 100 trees. The choice of Random Forest was because of the robustness of this technique to overfitting, its capabilities to capture non-linear relationships, and that they eliminate the need of the feature importance estimation as this is built in. It was a multidimensional and interpretable learner with gradient-boosted and SVM. The Random Forest classifier also proved to be rather stable during cross-validation, and it helped in achieving a better recall in the ensemble architecture.
- **Support Vector Machine (SVM):** The SVM model was set up with a radial basis function (RBF) kernel that would capture nonlinear relations between input variables and the outcome readmission. Since the classes

are not balanced, we have deployed cost-sensitive learning through a balanced class weight, where the instances of the minority classes will be more heavily penalized in misclassification. A good baseline was SVM because of its stability in high dimensions.

- **XGBoost:** The optimized gradient-boosting framework XGBoost was used since it has been successful in prediction tasks in healthcare. We used scale_pos_weight parameter to tune the effect of the majority and minority classes. The hyperparameters (learning rate, maximum depth, and number of estimators) were optimized by cross-validation. SHAP analysis was also helped by the natural feature importance of XGBoost.
- **LightGBM:** LightGBM was chosen due to its training speed and compatibility with large volumes of data. It was augmented with early stopping rules, to avoid overfitting, and optimized in experimental runs, both with grid search and optuna. We also trained focal loss with a similar purpose of making the algorithm less harsh on easy to classify majority examples and harsher on difficult to classify minority ones which led to modest improvements in minority class F1-score.
- **Stacked Ensemble Model:** We built a stacked ensemble to exploit the complementary powers of various learners. Base learners were XGBoost, LightGBM, Random Forest, and SVM. Their predictions belonged as input to a meta-learner Logistic Regression, which was selected because of its simplicity and interpretability. First, the Gradient Boosting classifier was used as a meta-model but was changed because of overfitting tendencies that were realized during validation. This combination continually performed better than single models in recall and generalization, and particularly following threshold tuning.

3.6 Threshold Tuning

Rather than using a default probability of 0.5 as a decision boundary, we adjusted the decision boundary to between 0.4 and 0.5 in steps of 0.05. The recall was best at 0.4 in the readmitted class (81 percent), and at 0.45, the trade-off between recall and precision was better (72 and 29 percent respectively) for the minority class. These thresholds were assessed in accordance with the F1-score of the minority class and clinical significance of false negatives minimization.

3.7 Evaluation Metrics

All models were assessed using:

- Accuracy (reported but de-emphasized)
- Precision, Recall, F1-score (for both major and minor class)
- Confusion matrices
- SHAP value summaries for feature contribution analysis
- Macro-averaged metrics

4 Results and Discussion

4.1 Performance Comparison of Models

The preliminary experiments also showed that accuracy is frequently used, but the results are deceptive in case of our imbalanced data. To cite but one example, the SVM model exhibited initial accuracy of 91% which however was entirely unsuccessful at detecting any patients belonging to the minority category (readmitted within 28 days). Following cost-sensitive training (the class_weight argument that takes the value of balanced), the recall of the positive class was boosted astoundingly to 72% and the F1-score to 0.42. XGBoost and LightGBM had better folds stability. After an optimal scale_pos_weight parameter tuning, the model obtained a macro F1-score of 0.71, and the recall up to 79 percent. Although LightGBM had great accuracy, it did not perform well in terms of recall - meaning that this model was more conservative at labeling patients at high risk. The Stacked Ensemble model was much better than individual models, especially in combination with threshold tuning and SHAP-based feature selection. At the threshold of 0.4, it gave a recall of 81 per cent, an F1-score of 0.40 in the minority class. It is, however, important to note that its performance was not biased by overrepresented false positives, which made the model the most clinically viable one.

The implementation code used for data preprocessing, and model training, evaluation is available at: https://github.com/aithasahith02/Predictive-analysis-of-Heart-Patients-Re-admission

4.2 Threshold Tuning Trade-offs for Minority Class

Table 1 Thresholds and Evaluation metrics

Threshold	Recall	Precision	F1-Score	Accuracy
0.50	69%	33%	0.45	85%
0.45	72%	29%	0.41	82%
0.40	81%	26%	0.40	78%

Table 1 summarizes how adjusting the decision threshold affects the model's performance on the minority class, clearly illustrating the trade-off between recall and precision. These results illustrate the core trade-off between recall and precision. A threshold of 0.4 maximized sensitivity, an important consideration for clinical decision support systems where missing a high-risk patient can have severe consequences.

4.3 SHAP Value Insights and Feature Importance

SHAP plots revealed that the most predictive features aligned well with known clinical markers of heart failure progression:

- NYHA Classification
- Killip Grade
- Consciousness at admission
- Systolic and Diastolic Blood Pressure
- Age Category
- Serum Creatinine

SHAP-based pruning also reduced overfitting, leading to more generalizable models and a 10% improvement in minority class F1 score across multiple learners.

4.4 Discussion

The general aim of the study was the creation of a prediction system of early readmission risk in heart failure patients that would be fair and accurate and based on the consideration of a problem with severe class imbalance coupled with feature redundancy and poor interpretability within classical ML pipelines. Our results confirm that being accurate is not enough, and particularly where the false negative outcomes are associated with risks to life. We were, however, able to make a model with clinically useful performance by setting the goal to maximize recall, using SHAP explanations and threshold tuning. In comparison with the MLMI 2022 project, where SMOTE and G-Mean based optimization were put in the foreground, our findings suggest that tuning with thresholds is a less complex yet flexible mechanism to boost the detection of the minority classes without adding synthetic points. Furthermore, our SHAP-assisted pruning algorithm has similar results to the results presented by Zahar et al. at ICAAI 2024 [7] and its findings confirm that explainable models can be performant, and it works particularly well to prune redundant clinical characteristics. Our benchmarked sampling methods include SMOTE, ADASYN, and SMOTE-Tomek. They are typical operations during unbalanced classification, and in this case gave mixed results, not to mention poorer performance. This is attributed to facts that in the literature the effects of synthetic sampling can be detrimental to generalization in some situations in that they occur when the distributions of features form among the classes overlap a lot. Our flexibility and stability were provided by the option of a stacked ensemble and a logistic regression meta-learner. The use of Gradient Boosted Trees as a meta-learner introduced overfitting, whereas the logistical regression allowed having a sharp decision boundary with no problematic overfitting. It is, however, important to note that the model was trained on a regional dataset of China yet the characteristics such as age, Killip grade, and

BP are standard and collected universally implying a possible generalization. Nevertheless, the next development will involve testing on the MIMIC-III [11] dataset to ensure that the model will be robust on diverse demographics and locations. While high recall minimizes missed high-risk patients, it increases false positives, potentially burdening clinicians with unnecessary follow-ups. To mitigate this, we propose integrating the model into EHR systems with tiered alerting and SHAP-based explanations, enabling clinicians to prioritize cases effectively. This approach balances sensitivity with workflow efficiency, reducing alert fatigue, and preserving trust in AI-assisted decision-making.

This study was performed using a regional dataset originated from hospitals in China, which may limit the generalizability of the findings to other patient populations and healthcare systems. Besides, neither external validation nor temporal validation were performed in the current work. As a result, the stability of the model and configurations may remain untested across institutions. It is also important to acknowledge the fact that a recall-oriented optimization may increase burden on clinicians due to the chance of an increase in number of false positives. So, the model is intended to work as a decision support system rather than a decision maker.

5 Conclusion

The current research introduces a machine learning framework including data preprocessing, feature selection based on SHAP, class imbalance resolution, and fine-tuning threshold in predicting early readmission of heart failure patients. The given stacked ensemble model has a high recall, better detection of minorities and actionable, interpretable, so it can be integrated into clinical practice. We feel that taking a step towards fairness-aware modeling and transparency of feature attribution, the work will have a positive impact on the implementation of AI in real-life hospital conditions, which is due to our departure with the traditional metrics concerning accuracy.

6 Future Research

Although our existing framework has good outcomes regarding the ability to predict early readmission risk of heart failure patients, its external validity has yet to be evaluated in external datasets. Next steps involve the validation of the model against the MIMIC-III database and other U.S-based EHRs to evaluate how generalizing the model acts against different populations and care settings. Potential and temporal validation are also another method we would like to consider providing reliability in the long term. Not only that, but we also seek to combine clinician-in-the-loop feedback as closer to actual deployment and threshold optimization. Should it be necessary, the generative AI approach, comprised of Variational Autoencoders (VAEs) or tabular GANs, could be considered to create synthetic samples of the minority category in extreme cases of imbalance, and guarantee finding clinically feasible solutions,

respectively. These methods may also be applicable in assisting balanced and fair training in cases where there is insufficient or biased actual data.

Acknowledgment

This research project was supported by the College of Education, Criminal Justice, Human Services, and Information Technology (CECH-IT) Graduate Student and Faculty Research Mentoring Grant from the University of Cincinnati (UC) in Ohio.

Contribution of each Author

Adib Zaman conceived the original project idea and provided overall mentorship and guidance throughout the research study. Sahith Aitha carried out background research, designed and performed multiple experiments for the problem statement. Sahith Aitha collected and preprocessed data, trained and evaluated the models, and prepared the initial manuscript draft. Adib Zaman reviewed the findings and work, providing critical feedback. Both the authors contributed to interpreting the results and approved the final manuscript.

Conflict of Interest Statement

The authors declare no conflict of interest.

References

1. Z. Zhang et al., “Explainable Machine Learning for Predicting 30-Day Readmission in Acute Heart Failure Patients,” iScience, vol. 27, no. 4, 2024. (Research article)
2. Y. Reddy, et al., “Readmissions in Heart Failure: It’s More Than Just the Medicine,” Mayo Clinic Proceedings, vol. 94, no. 10, pp. 1919–1921, 2019. (Journal Article)
3. Chen, S., Hu, W., Yang, Y., Cai, J., Luo, Y., Gong, L., Li, Y., Si, A., Zhang, Y., Liu, S., Mi, B., Pei, L., Zhao, Y., & Chen, F. (2023). Predicting Six-Month Re-Admission Risk in Heart Failure Patients Using Multiple Machine Learning Methods: A Study Based on the Chinese Heart Failure Population Database. Journal of Clinical Medicine, 12(3), Article 870. https://doi.org/10.3390/jcm12030870. (Research article)
4. Wu, J., Tao, G., Xie, S., Yang, H., Qi, F., Bao, N., Li, Z., Chang, G., & Xiao, H. (2025). Prediction of three-year all-cause mortality in patients with heart failure and atrial fibrillation using the CatBoost model. BMC cardiovascular disorders, 25(1), 466. https://doi.org/10.1186/s12872-025-04928-w. (Research article)

5. M.-Y. Yu and Y.-J. Son, “Machine Learning Models for Predicting 30-Day Heart Failure Readmission: A Systematic Review,” European Journal of Cardiovascular Nursing, vol. 23, no. 1, pp. 5–16, 2024. (Systematic Review Article)
6. A. Zahar et al., “Outcome Prediction Using SHAP and Ensemble Learning: A Clinical Case Study,” in Proc. Int’l Conf. on Artificial Intelligence (ICAAI), 2024. (Conference Paper)
7. A. Zahar, Z. Belkho, O. El Kalkha, F. Khennou, and O. El Meslouhi, “Detecting Cardiovascular Morbidity using MIMIC-III Database: Assessing Three Data Imbalance Handling Techniques,” in *Proc. ICAAI 2024*, London, United Kingdom, Oct. 2024, pp. 1–6. doi: 10.1145/3704137.3704162. (Conference Paper)
8. Zhang, Z., Cao, L., Zhao, Y., Xu, Z., Chen, R., Lv, L., & Xu, P. (2022). Hospitalized patients with heart failure: integrating electronic healthcare records and external outcome data (version 1.3). PhysioNet. https://doi.org/10.13026/5m60-vs44. (Research Dataset)
9. C. Yang, Y. Dong, J. Lu, and Z. Peng, “Solving Imbalanced Data in Credit Risk Prediction: A Comparison of Resampling Strategies for Different Machine Learning Classification Algorithms, Taking Threshold Tuning into Account,” in *Proc. MLMI 2022*, Hangzhou, China, Sept. 2022, pp. 1–11. doi: 10.1145/3568199.3568204. (Conference Paper)
10. L. Loezer, F. Enembreck, J. P. Barddal, and A. S. Britto Jr., “Cost-sensitive learning for imbalanced data streams,” in *Proc. 35th ACM/SIGAPP Symposium on Applied Computing (SAC)*, Brno, Czech Republic, 2020, pp. 4–10. doi: 10.1145/3341105.3373949. (Conference Paper)
11. Johnson, A., Pollard, T., & Mark, R. (2016). MIMIC-III Clinical Database (version 1.4). PhysioNet. https://doi.org/10.13026/C2XW26. (Research Dataset)

Sahith Aitha is an MLOps Engineer at Preston Ventures LLC, California, United States of America. He completed his master’s degree in information technology from the University of Cincinnati, Ohio, United States, graduating in 2025. His work and research interests focus on AI, Machine Learning, and applying computational models to solve real-world problems.

Dr. **Muhammad Adib Uz Zaman** is an assistant professor of IT at the University of Cincinnati, specializing in data science and health IT. Dr. Zaman utilizes machine learning and deep learning techniques to develop predictive models that assist physicians in critical decision-making processes. His research aims to optimize patient care and improve clinical outcomes.