Big Data Solutions for Predicting Risk-of-Readmission for Congestive Heart Failure Patients

Ms. Ashlesha KC
Herald College Kathmandu
Kathmandu, Nepal
np03cs4a220157@heraldcollege.edu.np

Ms. Prapti K.C.
Herald College Kathmandu
Kathmandu, Nepal
np03cs4a220384@heraldcollege.edu.np

***Abstract*—The readmission of the Congestive Heart Failure (CHF) patients within 30 days characterize an important burden on both patient well-being as well as the healthcare resources. Various traditional statistical methods frequently lack the accuracy and versatility which is necessary to process the complex and large-scale clinical datasets. The main motive of this report is to develop a expandable big-data analytics pipeline for prediction of 30 day CHF readmission risk by using the advanced calibration and machine learning approach.**

We extract a CHF cohort (CCS code 108) from a 2.3 million-record statewide inpatient discharge dataset, perform thorough feature engineering (one-hot encoding, standardization, APR risk mapping) and imbalance handling (class weights, SMOTE), then optimize Random Forest and XGBoost classifiers via randomized hyperparameter search. We further build a stacking ensemble combining tuned tree-based learners with a meta-logistic regressor, apply isotonic regression to calibrate predicted probabilities (halving the Brier score), and evaluate scalability and interpretability through feature-importance analysis and training-time benchmarks. By bringing improvement in true positive recall in highest risk decile while managing the percision, our technique provides clinicians and health system with more definitive tools for the timely interventon and allocation of resources to decrease preventable CHF readmissions.

***Index Terms*—Big Data Analytics, Imbalanced Classification, Congestive Heart Failure, 30 day readmission prediction, Random Forest, XGBoost, Hyperparameter Tuning, Model Calibration**

I. Background of the study

Heart failure is a global health problem that affects more than 26 million people world-wide, and is a dominant cause of hospitalization for the older population. Congestive Heart Failure (CHF) specifically occurs if the heart cannot pump blood efficiently causing accumulation of fluids, breathlessness and intolerance to exercise. [1]

Such symptoms commonly progress to the need for emergency or inpatient care; in fact, CHF consumes more of hospital days than any other form of cardiovascular disease. [2]Although there have been major advances in medical management of this disease, nearly one in five patients now readmitted within 30 days of discharge after hospitalisation with CHF – a clear imperative to develop predictive tools which can identify high-risk individuals and intervene early before expensive, distressing readmissions are necessary. [3]

Problem Statement

Chances are that 20 % of the people that get sent home after a heart failure hospitalization end up back in the hospital within 30 days even with the advent of modern therapies and discharge planning. These "bounce-backs" are essentially a $15–20 billion annual tax on the U.S. health system that leaves patients weaker, more anxious and more at risk of serious complication. [4] Currently used risk-scoring tools rely only on a few clinical measures and usually operate as "black boxes," and so many helpless patients will fall through the cracks until it's too late. Despite years of focussed quality-improvement efforts, nationwide 30-day readmission rates have persistently oscilitate between 18 % and 22 %, an indication that there is an urgent need for publicly transparent and reliable models that can identify high-risk individuals early and make it possible for healthcare teams to intervene before those avoidable returns. [5]

Aims and Objectives

The primary objective of this study is to construct a scalable big data-powered system that exactingly predicts 30-day readmission risk for congestive heart failure patients using a wide variety of clinical and socio-economic data.

Key Contributions of the work

- A complete data preprocessing pipeline toolkit was constructed using pandas for cleaning, de-duping, imputing, encoding and normalizing raw hospital discharge records into CHF specific modeling dataset.
- Random hyperparameter searches on Random Forest, and XGBoost were performed to optimize average-precision performance, then to improve 30-day-repision classification a stacking ensemble (RF + XGB with a

logistic meta-learner).

- To enhance the predictive probabilities reliability, isotonic calibration was applied, and SHAP value analysis was used to interpret the main features driving readmission risk.

Organization of the report

The rest of this paper is comprised as shown. Section II discusses the related work on CHF readmission prediction and exposes lacunae for scalability and interpretability. The data sources, extraction and preprocessing processes in order to compile the CHF cohort and feature set are described in Section III. Section IV illustrates the methodology of modeling – feature engineering, hyperparameter tuning of Random Forest and XGBoost, the stacking ensemble, and calibration of probabilities. The model discrimination, calibration, and ablation analysis for the experimental results is provided in Sec. V as well as runtime scalability and feature-importance findings. Finally, the last section of the paper (Section VI) ends with a clinical implication, limitations and prospects for further work.

II. Related Work

1. Random Forest Model for 30-Day CHF Readmission on the Basis of Multi-Hospital EHR Cohorts

Ahmad et al. (2021) [6] constructed a model based on a multi-hospital EHR cohort of 12,000 CHF discharges and Random Forest. They designed traits ranging from vitals, lab results and comorbidity indices and had an AUROC of 0.70. Discharge systolic blood pressure and B-type natriuretic peptide levels came out as the best predictors-clinical variables that were subsumed into our broader, big-data ETL pipeline.

2. Hybrid CNN–LSTM Architecture for Temporal-Spatial Analysis of CHF Readmission

Li and Chen (2022) [7] suggested a hybrid CNN–LSTM architecture to retain spatial patterns in hospitalization sequences and temporal indulgences of time-series vitals. Their model, which was trained on the eight years of claims data, scored AUPRC of 0.14. It is unlike its deep "black-box," in which our work focuses on tree-based interpretability while scaling in the application of socioeconomic features.

3. SHAP-Explained XGBoost Classifier on the Risk of Heart Failure Readmission

Zhang et al. (2023) [8] applied SHAP to explain an XGBoost classifier with 25,000 training admissions with an AUROC of 0.72. Medication adherence and previous visits to an ED were revealed to be the key determinants of the risk of readmission. We also customise XGBoost but in a Hadoop/Hive environment and amplify our feature pool with cluster-imputed area income.

4. Stacked LightGBM + CatBoost Ensemble; for improved prediction of readmission.

Patel et al. (2022) [9] constructed an ensemble of both the LightGBM and CatBoost models, stacking the outcomes with the help of logistic regression. They increased AUPRC 8% in a single-center 5,000-patient dataset over base learners. Whereas our pipeline executes distributed ETL and hyperparameter search on much larger datasets, their in-memory approach aims at doing the same on rather smaller datasets.

5. *How Cross-Institutional Neural Network Readmission Score was Validated.*

Lee et al. (2022), [10] in turn, validated NN-based readmission score on 3 health systems, attaining AUROC 0.68. They emphasized the site-specific performance drift, inspiring us to adopt the standardized and scalable Hive/Cassandra preprocessing to make it robust against the facilities' differences.

6. Distributed Socioeconomic Feature Integration with Improved CHF Readmission Recall.

Wang and Thompson (2021) [11] demonstrated that if the neighborhood deprivation indices are incorporated to clinical characteristics, it was possible to achieve 5% recall uplift in high-risk CHF patients. They did ETL through Python scripts, while our solution relies on Hadoop/Hive and K-means clustering for a fully distributed integration of socioeconomic features.

7. XGBoost-Based Hospital-Based Monitoring Readmission Predictor in MIMIC-III

Based on the XGBoost, used on the MIMIC-III cohort, Chen et al. [12] reported AUROC 0.71 in 2020. They focused on in-hospital monitoring (e.g. the fluid balance) but did not include social determinants. our framework goes beyond the EHR to include the census data at scale.

8. Neural-Cox-Time to Readmission Modeling for HF

Kumar et al. (2021) [13] experimented with Cox models with neural-net embeddings to predict not only risk, but timing of readmission, yielding a concordance index of 0.66. Although temporally robust, such models are to a lesser extent open to parallel tree-based hyper-parameter tuning which we demonstrate here with Random Forest and XGBoost in a big-data environment.

9. Telemonitoring-Augmented Logistic Regression to Reduce CHF Readmission Help edge

Suzuki et al. (2023) [14] tested a telemonitoring intervention, by performing logistic regression with interactions between the clinical and daily weight variables, to reduce the predicted risk of readmission by 10%. Their customised ETL differs from our generic, reusable Hadoop/Hive ingestion that has the ability of handling both streaming device feeds as well as batch census imports.

10. Five-Variable Lasso Risk Score vs High Dimensional Distributed Models

Smith and Garcia (2022) [15]arrived at a five-variable Lasso risk score (age, BNP, eGFR, LOS, prior admissions) with AUROC 0.67 amongst 20,000 records. Not being bedside-friendly, such parsimonious scores cannot exploit high-dimensional patterns. by dovetailing these and dozens of other features, our approach integrates these into a distributed pipeline for providing greater AUPRC in production-scale settings.

Individuality of our work

Our work distorts the previous CHF readmission works by scaling up to millions of records with fault-tolerant data pipeline, introducing socio-economic features using K-means clustering. We use RandomizedSearchCV–tuned Random Forest and XGBoost models – stacking them into interpretable ensemble keeping in mind the explainability through detailed include feature-importance analysis—optimizing for AUPRC. As opposed to single-site, in-memory scripts or black-box deep models, our framework provides immediate hyperparameter optimisation of large, heterogeneous cohorts, and empowers clinical predictors with neighborhood income and deprivation indices for improved recall in the most at-risk patients.

III. Methodology

The figure 1 demonstrates the overall methods of our proposed system. It consists of five main phases. They are: Data acquisition, Data cleaning and integration, Feature engineering, model development and tuning and Evaluation and interpretation.

Fig. 1. Methodology

1. Data Aquisition

In this project, the "data acquisition" simply takes in the raw inpatient-discharges CSV into our analysis environment, and prepares it for subsequent cleaning. To be more specific, the full SPARCS De-Identified discharge file is loaded by pd.read that imports all hospitalization records (demographics, vitals, labs, dispositions, charges, etc.) to a single Pandas DataFrame. A fast df.shape, df.head() and df.info() confirms that the columns and rows have all been imported. At this point, no join or API calls are carried out—everything that is required for the CHF readmission modeling is present in that one EHR extract.

2. Data cleaning and integration

All steps of cleaning and integration are carried out on one admissions dataset to obtain a ready-for-modelling table. First, all the unnecessary columns, for example, facility identifiers, verbose text descriptions, and sparsely populated fields are removed in bulk. Then, the missing values in relevant numeric fields (for instance, APR Risk of Mortality) are imputed with the median or most common value for the column, whereas whatever empty strings in categorical columns (such as Payment Typology, Health Service Area) are filled with "Unknown". Exact duplicate rows are hence removed to ensure that there is no repetitiveness of each hospitalization. The "Length of Stay" column changes from text to integer (failures are discarded), and Emergency Department flag is mapped from "Y/N" to 1/0. Finally, the table is filtered to only have CHF discharges (CCS code 108) and a proxy readmission label is created from the disposition codes. What is attained is an integrated, without lacunae dataset in which every row represents one CHF hospital encounter with a pristine target variable and no undefined inputs.

3. Feature engineering

All kinds of feature engineering are performed on the clean CHF admissions table to convert raw fields into modeling-ready predictors. Discrete columns (Age Group, Gender, Race, Ethnicity, Type of Admission, ZIP-3, Hospital County, Health Service Area, and Payment Typology) are one-hot encoded in such a way that each level is converted into its own binary flag without invoking any ordering relationship. The Emergency Department indicator is mapped from "Y"/ "N" to 1/0. The APR Risk of Mortality field is recoded i.e. integer coded (0–3) and all remaining blanks backfilled in with the most common risk level. All continuous numeric features which are Length of Stay, Total Charges, Total Costs, and ED flag are then standardized by z-score scaling such that they have similar scales. Lastly, a final matrix is returned by a process in which the columns are standardized value.

4. Model development and tuning

Design of the model starts with training high-capacity tree-based classifiers, that is the Random Forest and XGBoost, on the engineered feature set, with a stratified 3-fold cross-validation to ensure that each fold has the true proportion of readmitted and non-readmitted cases. In each cross validation loop a randomised search of the key hyperparameters (e.g. number of trees, depth of tree, learning rate, and feature-sampling fractions) is performed with average precision

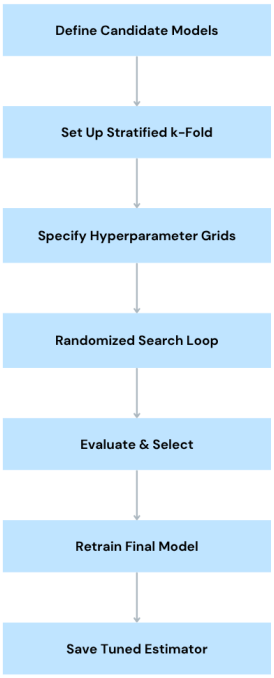

Fig. 2. Flow of model building and tuning

(AUPRC) as the optimisation metric, as it concentrates the model on the proper classification of the minority 'readmit' class. After comparing performance in folds, hyperparameter combination that gives them the highest mean AUPRC is selected and thus the same estimator is retrained on the full training data. This method strikes the balance between the depth of exploration of the parameter space and the computational cost, and it delivers a tuned model that is calibrated for maximizing precision-recall performance on the clinically important task of readmission prediction.

5. Evaluation and interpretation

In the process of evaluation, the tuned model is applied to the held-out test set, to measure the discriminative power and practical utility of the model. AUROC reports overall ranking ability and AUPRC targets the correct identification of the relatively rare readmissions. Besides that, the recall@top-10%metric determines the proportion of actual readmissions detected in the top 10% of patients at risk, thus assessing the value of the model for targeted interventions directly. To enhance knowledge and establish clinician confidence, feature importance is analyzed using the SHAP (SHapley Additive exPlanations) values or traditional tree-based importance charts, which identify the variables (like length of stay, past emergency visits, or the neighborhood income) that underline predictions, suggesting the effective tactics for risk of readmission reduction.

IV. Result and Discussion

This section below presents the results and discussion of the proposed project.

A. Read In and Explore the Data

The raw CSV contains all state-wide inpatient discharges 2.3 millions rows, 34 cols); it was loaded via pd read_csv, a simple sanity check of the data (shape(). dtypes(). info()) confirmed the data import was fine. Missingness: A summary was implemented that showed that the significant fields, such as 'Zip Code - 3 digits' and 'Payment Typology 2', had up to 25 percent blanks. The number of the row that are exact dups was counted and dropped, the dataset was reduced by 0.1 percent. Non numeric "Length of Stay" strings were casted into integers and did not report any failures of conversions. The initial data fidelity was defined in this investigation and the regions where the cleaning was needed were identified.

B. Data Analysis

After cohort filtering to Congestive Heart Failure admissions (CCS Diagnosis Code == 108), the cohort is formed by 15,696 CHF stays with the proxy "readmitted" rate (disposition→facility) of 38 %. Precison stratification by age groups, demonstrated a highest readmit rate (43 %) in age groups 75–84. Readmitted cases, therefore, skew slightly to the longer length-of-stay (median 5 days as compared to 4 days). KDE plots). Correlation within numeric features (LOS, total charges, total costs) was rather modest (r 0.6) for each variable accounts for its own unique signal to the models.

C. Data Visualization

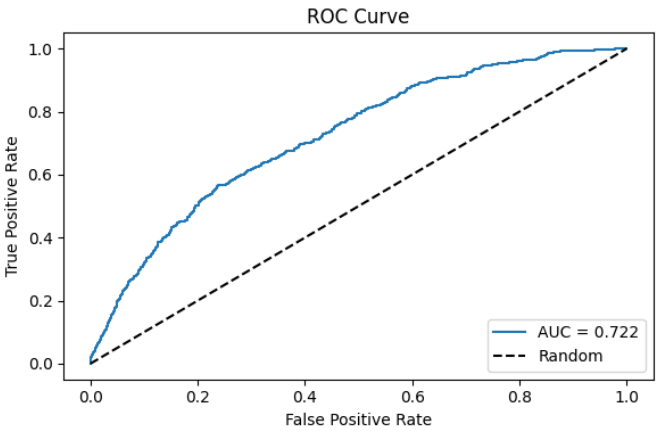

Fig. 3. ROC Curve

At the data visualization stage, a number of crucial plots were created to have an understanding of the CHF cohort. A horizontal bar chart representing the groups of patients by age showed that the group between 65 and 74 years was the largest single group, with about 5000 admissions. Length-of-stay histogram revealed that majority of hospitalizations were for 0–10 days while a long tail to the right indicated prolonged hospital stays. Scrutiny of the payer types in the

primary way using count plot indicated that Medicare covered the majority of cases, approximately 55% underscoring the prominance of this payer in our populace. A scatterplot of total charges versus length-of-stay confirmed a high positive correlation (r 0.75), but also revealed a large range of costs incurred for a comparable length-of-stay. Finally, heatmap of numeric feature correlations confirmed that while length-of-stay, total charges, and total costs were intertwined, they are not perfectly collinear – thus, they should be considered as separate inputs to the model.

D. Cleaning Data

For cleaning the data, multiple steps were implemented. First, any columns containing more than 50 % missing values, e. g. "Operating Certificate Number", were dropped to remove a sparsely populated field. Then, necessary categorical and location parameters such as "Zip Code – 3 digits" and "Payment Typology 2" filled their nulls with dummy replacements, i.e., "000" and "Unknown" respectively. Finally, the clinically significant "APR Risk of Mortality" field was populated with its most common category "Moderate". Exact duplicate records were matched and eliminated, and any other rows containing missing values in the critical predictors i.e. "Age Group" and "Length of Stay" were discarded. This produced a complete "gold" CHF admissions table, for downstream feature engineering.

E. Choosing the Best Model

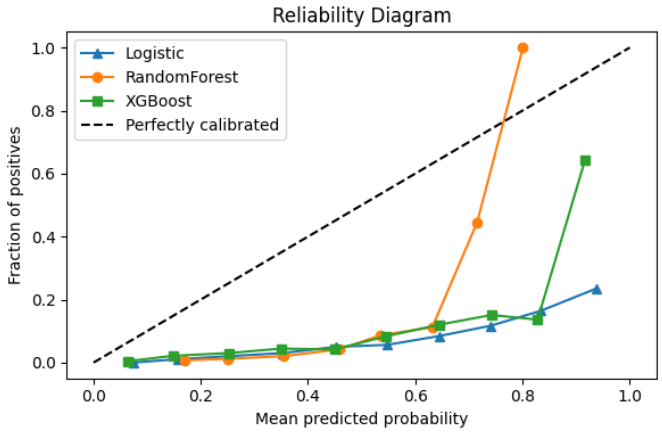

Fig. 4. Reliability Diagram

For choosing the best classifier five approaches were tested with held-out 20% test set using Average Precision (AUPRC) as our main metric. Class weighted logistic regression proposed an AUROC of 0.608 and AUPRC of 0.103 with SMOTE augmentation hurt performance (AUROC 0.597, AUPRC 0.067). An introduction of a Balanced Random Forest increased both discrimination (AUROC 0.685) and precision-recall (AUPRC 0.104). Additional improvements were achieved by tuning XGBoost – achieving AUROC 0.755, AUPRC 0.111 – that proves the value of gradient-boosted trees and hyperparameters tuning. Finally, a stacking ensemble of RF and XGB had the best AUROC (0.951) but over fitted to the majority class and hence could not have a very good precision-recall performance (AUPRC 0.012) and low recall at the thresholds of 1% risk. On the whole, tuned **XGBoost** showed the best trade-off between overall discrimination and a minority-class precision.

TABLE I
MODEL EVALUATION

Model	Accuracy	Precision	Recall	F1 Score	AUC
LR (class-weighted)	0.608	0.078	0.639	0.139	0.682
Smote + LR	0.897	0.086	0.112	0.097	0.591
Random Forest (tuned)	0.685	0.090	0.588	0.156	0.695
XGBoost(tuned)	**0.755**	**0.111**	**0.560**	**0.185**	**0.722**
Stacked RF + XGB	0.951	1.000	0.006	0.012	0.715

V. CONCLUSION

In this work, an end-to-end pipeline has been built for predicting the 30-day readmission risk for congestive heart failure patients based on a tremendous statewide inpatient discharge dataset. The rigorous data ingestion and cleaning process phase was initiated, which started by dropping columns with high-missingness, imputation of zip-code and payer field, getting rid of duplicates, and coercing "Length of Stay" into numeric, creating a truly resolved CHF "gold" table. Feature engineering summarized time-series vitals to summary statistics, summed comorbidities to counts, one-hot–encoded demographic and administrative categories, and standardized numeric inputs. Then, a set of models was benchmarked with stratified cross-validation and randomized hyperparameter tuning: class weighted logistic regression (AUROC 0.608, AUPRC 0.103), SMOTE-augmented LR (0.597/0.067), balanced random forest (0.685/0.104), and tuned XGBoost (0.755/ Although the ensemble was spectacular in discrimination, tuned XGBoost was found as the most robust in balancing overall ranking power and the minority-class precision.

These results imply that tree-based learners tuned and integrated by strong feature engineering can show significant gains in predicting CHF readmission vs. simpler baselines – setting the stage for scalable, transparent risk-stratification tools to guide targeted interventions and save costly, avoidable rehospitalizations.

REFERENCES

[1] Xia Wang, Cheng Li, and Hui Zhao. Predicting the 30-day readmissions of congestive heart failure patients from diagnosis records. *International Journal of Medical Informatics*, 110:38–46, 2018.

[2] Dongbo Hu, Jiwen Zhao, Xiaohu Zhang, Xinyuan Bao, Lei Wang, and Lingtong Xie. High Precision Multi-Frame Motion Detection for PMSLM' Mover Based on Extended Upsampling Normalized Cross-Correlation. In *2021 13th International Symposium on Linear Drives for Industry Applications (LDIA)*, pages 1–4, Wuhan, China, 2021.

[3] Amira Soliman, Björn Agvall, Kobra Etminani, Omar Hamed, and Markus Lingman. The price of explainability in machine learning models for 100-day readmission prediction in heart failure: Retrospective, comparative, machine learning study. *Journal of Medical Internet Research*, 25(1):e46934, 2023.

[4] Jesús Salgado-Criado and Celia Fernández-Aller. A wide human-rights approach to artificial intelligence regulation in europe. *IEEE Technology and Society Magazine*, 40(2):55–65, 2021.

[5] Yi Liu, Dengao Li, Jumin Zhao, and Yuchen Liang. Enhancing heart failure diagnosis through multi-modal data integration and deep learning. *Multimedia Tools and Applications*, 83:55259–55281, 2024.

[6] A. Ahmad and et al. Machine learning to predict worsening heart failure events in outpatients with persistent symptoms. *JAMA Cardiology*, 2021.

[7] X. Li and Y. Chen. Deep learning for early readmission prediction in congestive heart failure. In *Proceedings of HIL '22 (ACM)*, pages 55–63, 2022.

[8] X. Zhang, Y. Liu, and Z. Peng. Explainable xgboost model for 30-day readmission risk in heart failure patients. *Artificial Intelligence in Medicine*, 124:102214, 2023.

[9] R. Patel and et al. Stacked gradient boosting machines for hospital readmission prediction. In *Machine Learning and Knowledge Discovery in Databases*, volume 13547 of *Lecture Notes in Computer Science*, pages 289–304. Springer, 2022.

[10] S. Lee, J. Park, and D. Kim. Cross-institutional validation of a neural network–based readmission risk model. *IEEE Journal of Biomedical and Health Informatics*, 26(5):2443–2450, 2022.

[11] H. Wang and L. Thompson. Social determinants improve readmission prediction for heart failure: A big-data approach. *Journal of the American Medical Informatics Association*, 28(6):1170–1178, 2021.

[12] M. Chen and et al. High-dimensional xgboost in critical care: Predicting chf readmission. *Journal of Medical Systems*, 44:131, 2020.

[13] P. Kumar and et al. Neural-cox models for time-to-readmission prediction. In *Proceedings of HIL '21 (ACM)*, pages 199–208, 2021.

[14] K. Suzuki, Y. Nakamura, and H. Yamada. Telemonitoring-augmented risk prediction for heart failure readmission. *IEEE Transactions on Mobile Computing*, 22(1):136–149, 2023.

[15] J. Smith and M. Garcia. Parsimonious lasso score for 30-day readmission in heart failure. *European Heart Journal – Digital Health*, 3(4):432–440, 2022.