Evaluating the Trade-off Between Predictive Performance and Model Fairness in Sudden Cardiac Death Risk Stratification

H. Ivandic*, B. Pervan*, J. Knezovic*, M. Puljevic†‡, A. Jovic*
* University of Zagreb, Faculty of Electrical Engineering and Computing, Zagreb, Croatia
† University Hospital Centre Zagreb, Zagreb, Croatia
‡ University of Zagreb, School of Medicine, Zagreb, Croatia
hana.ivandic@fer.unizg.hr

***Abstract*—Predicting implantable cardioverter-defibrillator (ICD) activation using machine learning (ML) is a promising approach for identifying patients at high risk of sudden cardiac death (SCD). Building on our previous optimization of SCD risk stratification ML models, this study investigates how data preprocessing affects model performance, feature importance, and fairness. We evaluate whether the native handling of missing data by gradient boosting algorithms, such as CatBoost (CB), outperforms traditional manual preprocessing and sampling techniques. Our analysis compares two pipelines: (1) multiple ML models trained on numerically encoded, manually imputed, scaled, and sampled data, and (2) a CB model trained on the original, unprocessed dataset, leveraging its native ability to handle missing values and categorical data. Using SHAP for explainability and fairness metrics to assess potential performance disparities, we quantify the impact of preprocessing techniques on the models' ability to identify high-risk patients across different groups. Results indicate that manual preprocessing negatively impacts CB performance. Furthermore, we confirm enhanced predictive performance on minority groups, specifically female patients, younger individuals, and those with LVEF > 35%.**

***Keywords*—machine learning, SHAP, fairness, sudden cardiac death, implantable cardioverter-defibrillator**

I. Introduction

Sudden cardiac death (SCD) is an unexpected natural death caused by a sudden heart rhythm malfunction, preventing the delivery of oxygenated blood to the rest of the body. Currently, the primary prevention of SCD relies on implantable cardioverter-defibrillators (ICDs), devices that reduce arrhythmic mortality by monitoring the heart rhythm and delivering an electric shock in case of heart malfunction. Despite the existence of such devices, SCD is still one of the leading causes of cardiovascular mortality [1], [2]. The criteria for selecting patients eligible for ICD implantation are defined by the European Society of Cardiology (ESC) guidelines [3] and, while supporting the ICD implantation for patients at increased arrhythmic risk, they mostly rely on the Left Ventricular Ejection Fraction (LVEF). The ICD implantation is recommended if the LVEF value, the percentage of blood pumped from the left ventricle with each heartbeat, is $\leq 35\%$.

Despite the existing guidelines, many patients who receive an implant never suffer an arrhythmia that triggers the device activation. On the other hand, many patients who do not meet the ESC criteria for ICD implantation fall victim to SCD. These facts highlight the potential of machine learning (ML) models to refine patient selection by predicting the individualized risks of SCD.

In medical ML applications, it is vital to ensure algorithmic fairness, particularly when dealing with imbalanced datasets. It is essential to ensure that the models provide equal opportunity for treatment, in this case, an ICD implant, regardless of the patient's demographic characteristics. Although many studies focus on optimizing the ML models to stratify the SCD risk, not many investigate the fairness of their models.

Rajkomar et al. [4] highlighted the importance of ensuring fairness in medical applications of ML models. They described the possible causes for model bias and listed a set of recommendations to minimize healthcare disparities. Additionally, multiple studies [5], [6] emphasized that missing values can cause ML model unfairness. Reiner et al. [7] performed a statistical analysis of sudden cardiac arrest stratified by the patients' race. On the other hand, Chen et al. [8] compared the performance of a Logistic regression (LR) model in predicting hospital mortality from medical notes with respect to patient race, gender, and insurance type, and predicted psychiatric readmission within 30 days from psychiatric notes with respect to patient insurance policy. Kabir et al. [9] reviewed the challenges of fair model development, especially focusing on SMOTE-driven techniques. They reported a notable improvement in classification performance after using SMOTE oversampling. A study by Chen et al. [10] proposed methods for dealing with missing, imbalanced, and sparse features specifically for SCD prediction models.

This study investigates the algorithmic fairness of several ML models, with a particular emphasis on the impact of data preprocessing pipelines, specifically missing value imputation and sampling techniques. Furthermore, we investigate the trade-off between predictive performance and fairness. Finally, the analysis explores how the prepro-

TABLE I: Age groups.

Label	Age range
1	< 45
2	45 - 54
3	55 - 64
4	$\geq$ 65

cessing strategies alter the underlying feature importance of the predictive models.

The rest of the paper is structured as follows. In Section II, we first introduce the dataset used in the study. Then, in Section III, we describe the methodology for model training, fairness evaluation, and feature importance analysis. In Section IV, we present the performance of different models on several patient groups, and in Section V, we analyze the feature importance. Section VI discusses the results of the study, while Section VII concludes the paper.

II. Dataset

The dataset used in the study contains demographic, clinical, laboratory, and device-derived data of patients who underwent ICD implantation at a Croatian tertiary center. It was first introduced and analyzed by Puljevic et al. [11]. The same dataset was later used by Ivandic et al. [12], [13], who applied and compared the performance of several ML models for ICD activation prediction, as the dataset also contains information about the device activation. The preprocessing steps were described in detail in [13], and the final dataset consisted of 607 patient records. Apart from anonymization, they included numerical encoding, feature selection and adjustment, and missing value imputation. The current study followed the same design for most of the models, except for the CatBoost algorithm, for which numerical encoding and missing value imputation were omitted.

Although the dataset includes 32 features in addition to patient ID, activation type, and activation appropriateness, this study focused on the fairness of the models across three features: patient gender, age group, and LVEF. These features were selected for further analysis as the patient gender and age are the only available demographic variables, and the LVEF is, according to the ESC guidelines [3], one of the main criteria for selecting patients for ICD implantation.

While the patient's gender is a categorical feature, their age and LVEF value are discrete. Thus, the patients were divided by age into four age groups, as defined in Table I. On the other hand, patients were divided by the LVEF value into two groups, with one group having the LVEF $\leq$ 35 and the other having the LVEF $>$ 35. This grouping was chosen because the 35% is the LVEF threshold defined by the ESC guidelines [3] as the main selector for the ICD implantation.

The distributions of the selected features in the dataset are presented in Figure 1. As displayed in the charts, the groups were not equally represented. There were more male than female patients, and the older age groups were far more represented than the younger ones, with only 11.9% of patients younger than 45. Thus, the age groups < 25, 25 – 34, and 35 – 44 usually used in research were combined into a single group. A positive trend was observed between patient age and cohort size, with the largest group comprising the oldest patients. With the LVEF as the main criterion for ICD implantation, the observed higher percentage of patients with the LVEF value $\leq$ 35 was expected.

It is also important to emphasize that the appropriate ICD activation occurred in only 27.5% of patients, confirming that the current patient selection criteria yield a high number of false positives. Therefore, ML models may offer potential for refining patient selection and reducing the number of false positives while preserving high recall.

III. Methodology

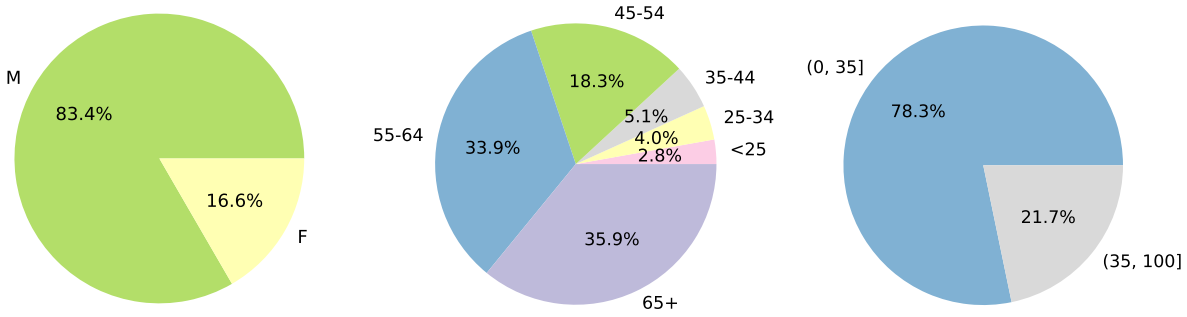
In this study, we address three key research questions:

1) How do model performance and feature importance change with and without the preprocessing steps?
2) How do different models perform across various patient groups?
3) How does the preprocessing influence model fairness?

The experiments were organized into two pipelines described in Table II. The first pipeline was described in [13]. After the missing value imputation, feature scaling, and test set sampling, the models were trained on all available features using a 10-fold cross-validation grid search. Models tested in this scenario include the CatBoost (CB), Logistic regression (LR), and Multinomial Naive Bayes (MNB) models. SMOTE (Synthetic Minority Oversampling Technique) [14] sampling was used for CB and MNB, while for the LR model, the dataset was subsampled as reported in the previous study [13].

The second pipeline was applied only to the CB model, as it can handle missing values natively. In this case, the model was trained on the original, not numerically encoded, features without missing value imputation, scaling, or sampling. To handle the class imbalance, more weight was put on the samples with ICD activation.

These pipelines were selected to inspect the impact of feature preprocessing and sampling techniques on model fairness. In both pipelines, the dataset was split into a train and test set in an 80:20 ratio, and the models were trained to optimize the F2-score. This metric was chosen because it considers both recall and precision, but assigns higher importance to recall, a critical requirement in medical applications. The fairness of the models was evaluated by comparing their performance on the predefined groups using the Python *fairlearn* library [15]. The ratios for precision (PPV), recall (true positive rate, TPR), and false positive rate (FPR) were calculated across the defined patient subgroups to identify potential disparities. The



(a) Patient gender (b) Patient age (c) Patient LVEF

Fig. 1: Distribution of the observed features.

TABLE II: Model training pipeline steps.

Step	Pipeline 1	Pipeline 2
Feature numerical encoding	+	-
Missing value imputation	+	-
Scaling	+	-
Sampling	+	-
Optimization metric	F2-score	F2-score
Evaluation	10-fold cross-validation	10-fold cross-validation
Hyperparameter selection	Grid search	Grid search
Trained models	CB, LR, MNB	CB

group with the highest F2-score was designated as the reference (denominator) for each observed feature. Consequently, female patients served as the reference group for patient gender, while patients with LVEF $> 35\%$ were the reference for the LVEF feature. For patient age, which comprised more than two subgroups, the < 45 age group was selected as the referent. Finally, the fairness ratios were calculated as follows:

$$PPV\_ratio = \frac{PPV_{observed}}{PPV_{referent}} \quad (1)$$

$$TPR\_ratio = \frac{TPR_{observed}}{TPR_{referent}} \quad (2)$$

$$FPR\_ratio = \frac{FPR_{observed}}{FPR_{referent}} \quad (3)$$

SHAP (Shapley additive explanations) analysis designed by Lundberg and Lee [16] was applied to the CB models to explain the impact of different features. The analysis was applied to the model that was trained on the dataset that used numerical encoding of the features, value imputation, scaling, and SMOTE sampling, and also to the model that was trained on the original dataset to assess the impact of the data preparation steps on the model decisions.

TABLE III: Overall model performance.

Model	Pipeline	ACC[1]	PPV	TPR	FPR	F2
CB	2	0.51	0.36	0.94	0.66	0.71
CB	1	0.39	0.31	0.97	0.83	0.68
LR [12], [13]	1	0.65	0.46	0.86	0.45	0.73
MNB [13]	1	0.66	0.47	0.78	0.39	0.69

[1] Accuracy (ACC)

IV. Results

The results are organized into three subsections, with each analyzing the models' performance grouped by different features. Subsection IV-A compares the model performance based on the patient gender, while Subsections IV-B and IV-C compare the model performance according to the patient age and LVEF value, respectively. Each table in the following subsections presents the metric ratios defined in Section III for different ML models. In addition to the models and their respective ratios, the tables also specify the pipeline used for each configuration.

Table III summarizes the overall results of the observed models. The results for the LR and the MNB were already published in [13]. As shown in the table, the LR model achieved the highest F2-score, while the CB model achieved the highest recall, reaching 94% on the original dataset and 97% when combined with preprocessing steps. Despite the highest recall, the CB models, especially with data preprocessing, had the poorest performance regarding the remaining metrics.

The impact of the preprocessing was not uniform across the evaluated models. While it was essential for the performance of the LR and MNB models, it led to degradation in the CB model's performance. Since CB is designed to handle raw data natively, manual preprocessing likely caused information loss and introduced noise to the data, resulting in a drop in accuracy and precision and an increased FPR.

The LR model with preprocessing steps emerged as the optimal configuration. It achieved a high TPR while maintaining a relatively low FPR. Although CB models showed

TABLE IV: Model fairness ratios for patient gender (male group / female group).

Model	Pipeline	PPV	TPR	FPR
CB	2	0.77	1.09	1.12
CB	1	0.69	0.96	1.16
LR	1	0.90	1.01	1.21
MNB	1	0.92	1.13	1.32

higher TPR, their high FPR suggests low discriminative power, indicating that they achieved high sensitivity by over-predicting the positive class.

It is important to emphasize that lower precision was expected as the models were trained to prioritize recall. According to the current guidelines, all patients were considered high risk (positive class), so the primary goal of the models was to minimize false negatives to ensure safety, while secondary efforts focused on reducing the false positive rate.

A. Patient gender

Table IV presents model metric ratios by patient gender. The referent metric values were derived from the model performance achieved on the group of female patients, since the F2-score was higher for this group across the majority of models. All models achieved a higher PPV and a lower FPR for female patients. The LR and MNB models demonstrated superior fairness in terms of PPV, and the CB models in terms of FPR. The CB model that included preprocessing was the only one to achieve a lower TPR for male patients. Preprocessing increased the fairness disparity within the PPV and FPR metrics. Although the TPR disparity remained marginal, the performance trend shifted from favoring male patients to favoring female patients.

Considering the fact that in medical applications the primary objective is the identification of all high-risk patients, TPR is considered the most critical metric. While performance disparities in PPV and FPR favored female patients, a higher TPR was observed for male patients in three of the four evaluated models. Clinically, this suggests that the models tended to predict the positive class more frequently for male patients. While this resulted in an increased FPR and a drop in PPV, it ensured a higher number of recognized true positive patients. Ultimately, the LR model emerged as the most unbiased, offering an optimal balance of metric ratios. It achieved high parity in TPR while maintaining acceptable ratios for PPV and FPR.

B. Patient age

The performance disparities across different patient age groups are shown in Table V. In this case, group 1 with patients younger than 45 was selected as the reference. Analysis of the metric ratios revealed consistent performance variations favoring the reference group across all evaluated models. These disparities were most pronounced in PPV and FPR, whereas TPR exhibited a higher degree of parity. Specifically, the highest discrepancy in PPV was observed for group 4 across all models. Regarding TPR, the CB models also showed the largest disparity for group 4, while the LR and MNB models did so for group 3.

TABLE V: Model fairness ratios for patient age group (age group 2, 3 or 4 / age group 1).

Model	Pipeline	Group	PPV	TPR	FPR
CB	2	2	0.68	1.00	0.72
		3	0.73	1.00	0.50
		4	0.30	0.75	0.71
CB	1	2	0.59	1.00	1.00
		3	0.63	1.00	0.71
		4	0.38	0.83	0.82
LR	1	2	0.53	0.86	2.16
		3	0.49	0.77	1.52
		4	0.44	0.91	2.16
MNB	1	2	0.58	0.71	1.52
		3	0.49	0.62	1.20
		4	0.44	0.91	2.08

TABLE VI: Model fairness ratios for patient LVEF value (LVEF $\leq$ 35 group / LVEF $>$ 35% group).

Model	Pipeline	PPV	TPR	FPR
CB	2	0.87	0.91	0.66
CB	1	0.76	0.96	0.84
LR	1	0.68	0.79	0.53
MNB	1	0.72	0.67	0.39

Both the LR and MNB models demonstrated significantly larger differences in FPR than the CB models, with the number of false positives doubling relative to the reference group in certain instances. For the CB model, the preprocessing steps improved group parity regarding TPR and FPR. However, PPV disparities remained substantial. While the disparity slightly decreased for group 4, it increased for groups 2 and 3.

C. LVEF value

Finally, Table VI displays the model metric ratios stratified by the LVEF values, with patients having LVEF $>$ 35% serving as the reference group. Regarding this feature, all models exhibited a consistent performance disparity favoring the reference group in terms of PPV and TPR. Conversely, a lower FPR was observed for patients with LVEF $\leq$ 35%. Similar to the results observed for patient age, the CB models demonstrated greater parity than the other two evaluated models. While for the CB models the preprocessing successfully reduced disparity in TPR and FPR, the disparity favoring patients with LVEF $>$ 35% increased in terms of PPV.

V. Explainability

The models' reasoning was analyzed using SHAP values. In our previous study [13], the LR and MNB models identified several features as the most influential: (1) VT frequency, (2) patient follow-up time, and (3) serum creatinine level. Higher VT frequency and longer follow-up time increased the likelihood of predicting an ICD activation. In this study, we repeated the same experiment for the CB model to investigate how the missing value imputation influenced the feature importance. The SHAP value graphs are presented in Figure 2. The CB model, trained the same way as the LR and MNB models using preprocessing, confirmed the influence of VT frequency and serum creatinine level. It also highlighted preimplantation ventricular tachycardia as a risk enhancer, while diuretic medication, low ejection fraction as an indication for ICD implantation, and preimplantation atrial fibrillation (AF)/atrial undulation (AU) were associated with lower risk estimates.

The analysis of the CB trained on the original dataset without preprocessing also confirmed the importance of VT frequency. Missing values reduced the likelihood of ICD activation prediction, while higher values of the VT frequency increased the probability of the model predicting an SCD. This was an expected observation as the VT frequency was missing in cases where the patients did not suffer a ventricular tachycardia. Opposite to the other CB model, the preimplantation ventricular tachycardia in the ECG decreased the chances of predicting a positive class, while low ejection fraction as an indication for ICD implantation increased it. While the feature importance was more transparent in the CB model, which utilizes imputation and sampling, this configuration resulted in diminished predictive performance. Consequently, it can be inferred that the remaining features, in this specific setup, lack sufficient discriminative power to maintain high performance.

VI. Discussion

The main aim of the study was to examine the influence of data preprocessing on the model performance, feature importance, and fairness. Specifically, the goal was to investigate the key research question defined in Section III. The first question focused on the performance of the models and feature importance with and without the preprocessing steps. The performance was also compared to other models reported in our previous studies [12], [13]. The drop in ACC, PPV, and F2-score, and the increase in FPR confirmed the degradation of the CB model after preprocessing. Compared to the LR and MNB models, both versions of the CB achieved higher TPR but lower ACC and PPV, and higher FPR. While the CB model without preprocessing outperformed the MNB in terms of F2-score, it fell short of the LR model, which remained the top-performing approach.

In both CB models, VT frequency emerged as one of the most important features, with high values favoring the probability of predicting a positive class. Nevertheless, the influence of several features, such as preimplantation VT in ECG and EF as an indication for device implantation, was opposite in the two versions of the CB models. Additionally, features with a high rate of missing values that were imputed within Pipeline 1 were emphasized as some of the most influential. These include the device manufacturing company, VT frequency, NYHA functional class, and diuretic medication, which could explain the drop in the model performance when using preprocessing. These results showed that the value imputation introduced noise to the data, resulting in performance degradation.

The second question investigated the model performance across several patient groups stratified by patient gender, age, and the LVEF value. In the case of patient gender, a disparity favoring female patients can be observed in PPV and FPR, while most models favored male patients in terms of TPR. The performance difference was more emphasized in CB models. As for the patient gender, all models consistently favored patients younger than 45. For the CB models, the lowest performance was shown for age group 4, while the LR and MNB models demonstrated the lowest TPR for age group 3. In terms of LVEF, the disparity favoring the group with LVEF $> 35\%$ was consistent across all models but more emphasized for the LR and MNB classifiers.

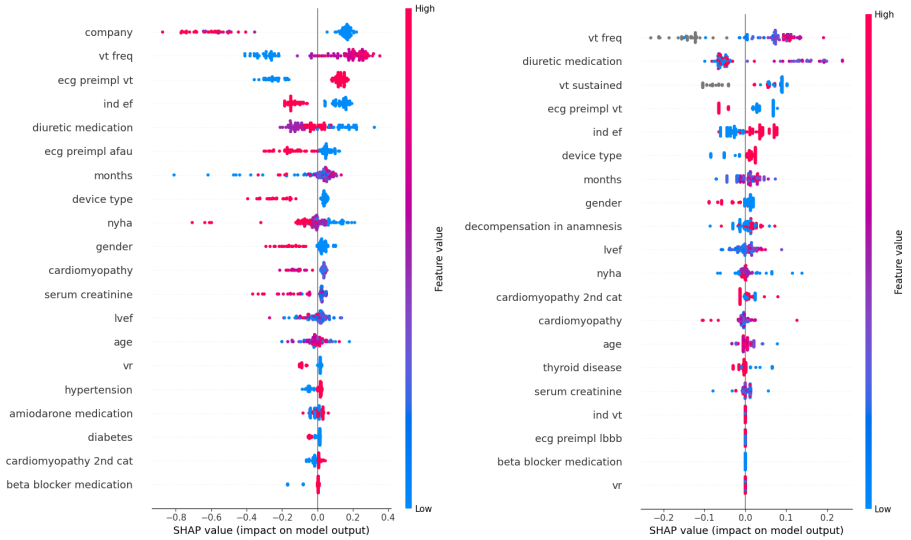
The last question dealt with the influence of the preprocessing steps on the fairness of the CB model. In the case of patient gender, the disparity between the groups increased. The preprocessing increased the disparity in PPV for age groups 2 and 3, while it decreased it for age group 4 and all groups in TPR and FPR. Regarding the LVEF value, similar to patient age, the preprocessing increased the disparity in terms of PPV, while it decreased it in TPR and FPR.

Regarding the trade-off between predictive performance and model fairness, the LR model exhibited the optimal balance between avoiding overdiagnosis and ensuring positive samples were not missed. Although it achieved the highest parity in patient gender, it fell short of the CB model results for patient age group and the LVEF values.

VII. Conclusion

This paper evaluates the performance and fairness of LR, MNB, and CB models for SCD risk stratification, specifically assessing the impact of preprocessing on the CB model. Our results showed that preprocessing degraded the discriminative power of the CB model. Despite the significant class imbalance within the dataset, our findings reveal a paradoxical performance trend. The models demonstrated superior predictive performance for smaller groups, specifically female patients, younger patients, and patients with LVEF > 35, which might be attributed to a lower intra-group variance of the smaller groups compared to the larger, more heterogeneous groups.

Several limitations of the study warrant acknowledgment. First, the relatively small and imbalanced dataset



(a) Pipeline 1 (b) Pipeline 2

Fig. 2: SHAP values for the CB models trained using the two pipelines described in Table II.

may constrain the models' generalizability. Consequently, validation on larger external cohorts is essential to confirm the findings of the study. Furthermore, in alignment with the scope of this study, the analysis was restricted to three variables. While patient gender and age served as standard demographic variables for bias assessment, LVEF was also included as the primary clinical criterion for patient selection. Future studies will examine additional features to further evaluate potential performance disparity across the models. Additionally, future work will include model fairness assessment after using fairness-aware algorithms.

REFERENCES

[1] J.-P. Empana, I. Lerner, E. Valentin, F. Folke, B. Böttiger, G. Gislason, M. Jonsson, M. Ringh, F. Beganton, W. Bougouin, E. Marijon, M. Blom, H. Tan, and X. Jouven, "Incidence of Sudden Cardiac Death in the European Union," *JACC*, vol. 79, no. 18, pp. 1818–1827, 2022.

[2] M. Zuin, S. Mohanty, R. Aggarwal, M. Bertini, B. Bikdeli, N. Hamade, H. Leyva, A. Natale, G. Boriani, and G. Piazza, "Trends in Sudden Cardiac Death Among Adults Aged 25 to 44 Years in the United States: An Analysis of 2 Large US Databases," *Journal of the American Heart Association*, vol. 14, no. 1, p. e035722, 2025.

[3] K. Zeppenfeld, J. Tfelt-Hansen, M. D. Riva, B. G. Winkel, E. R. Behr, N. A. Blom, P. Charron, D. Corrado, N. Dagres, C. D. Chillou *et al.*, "2022 ESC Guidelines for the management of patients with ventricular arrhythmias and the prevention of sudden cardiac death: Developed by the task force for the management of patients with ventricular arrhythmias and the prevention of sudden cardiac death of the European Society of Cardiology (ESC) Endorsed by the Association for European Paediatric and Congenital Cardiology (AEPC)," *European Heart Journal*, vol. 43, no. 40, pp. 3997–4126, 10 2022.

[4] A. Rajkomar, M. Hardt, M. Howell, G. Corrado, and M. Chin, "Ensuring Fairness in Machine Learning to Advance Health Equity," *Annals of Internal Medicine*, vol. 169, pp. 866–872, 12 2018.

[5] Y. Zhang and Q. Long, "Assessing Fairness in the Presence of Missing Data," in *Proceedings of the 35th International Conference on Neural Information Processing Systems*, ser. NIPS '21. Red Hook, NY, USA: Curran Associates Inc., 2021.

[6] E. Getzen, L. Ungar, D. Mowery, X. Jiang, and Q. Long, "Mining for equitable health: Assessing the impact of missing data in electronic health records," *Journal of Biomedical Informatics*, vol. 139, p. 104269, 2023.

[7] K. Reinier, A. Sargsyan, H. Chugh, K. Nakamura, A. Evanado, D. Klebe, R. Kaplan, K. Hadduck, D. Shepherd, C. Young, A. Salvucci, and S. Chugh, "Evaluation of Sudden Cardiac Arrest by Race/Ethnicity Among Residents of Ventura County, California, 2015-2020," *JAMA Network Open*, vol. 4, no. 7, pp. e2 118 537–e2 118 537, 07 2021.

[8] I. Y. Chen, P. Szolovits, and M. Ghassemi, "Can AI Help Reduce Disparities in General Medical and Mental Health Care?" *AMA Journal of Ethics*, vol. 21, pp. E167–179, 02 2019.

[9] M. A. Kabir, M. Ahmed, S. Begum, S. Barua, and M. R. Islam, "Balancing Fairness: Unveiling the Potential of SMOTE-Driven Oversampling in AI Model Enhancement," in *Proceedings of the 2024 9th International Conference on Machine Learning Technologies*, ser. ICMLT '24. New York, NY, USA: Association for Computing Machinery, 2024, p. 21–29.

[10] X. Chen, H. Chen, S. Nan, X. Kong, H. Duan, and H. Zhu, "Dealing With Missing, Imbalanced, and Sparse Features During the Development of a Prediction Model for Sudden Death Using Emergency Medicine Data: Machine Learning Approach," *JMIR Medical Informatics*, vol. 11, 01 2023.

[11] M. Puljevic, E. Ciglenecki, V. Pasara, I. Prepolec, M. D. Dosen, P. Hrabac, A.-M. Brekalo, M. L. Bencic, M. Krpan, R. Matasic, B. Pezo-Nikolic, D. Puljevic, D. Milicic, and V. Velagic, "CRO-INSIGHT: Utilization of Implantable Cardioverter Defibrillators in Non-ischemic and Ischemic Cardiomyopathy in a Single Croatian Tertiary Hospital Centre," *Rev Cardiovasc Med.*, vol. 26, 2025.

[12] H. Ivandic, B. Pervan, V. Velagic, A. Jovic, and M. Puljevic, "Improving Risk Stratification in Sudden Cardiac Death Using Interpretable Machine Learning: A Clinical Perspective," *Healthcare*, vol. 13, no. 21, p. 2788, 11 2025.

[13] H. Ivandic, B. Pervan, M. Puljevic, V. Velagic, and A. Jovic, "Machine Learning-Based Risk Stratification for Sudden Cardiac Death Using Clinical and Device-Derived Data," *Sensors*, vol. 26, no. 1, 2026.

[14] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," *Journal of Artificial Intelligence Research*, vol. 16, no. 1, p. 321–357, 6 2002.

[15] S. Bird, M. Dudík, R. Edgar, B. Horn, R. Lutz, V. Milan, M. Sameki, H. Wallach, and K. Walker, "Fairlearn: A toolkit for assessing and improving fairness in AI," Microsoft, Tech. Rep. MSR-TR-2020-32, 5 2020.

[16] S. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, ser. NIPS'17. Red Hook, NY, USA: Curran Associates Inc., 2017, p. 4768–4777.