FAMEWS: a Fairness Auditing tool for Medical Early-Warning Systems Reproduction Project

Fahad Zia

Abstract

The objective for this project is to reproduce the results of a selected research paper. This project focuses on FAMEWS, an auditing tool for medical early warning systems (EWS). The study investigates the correlation between accuracy and patient demographics including age, sex, race, and type of admission within early warning systems. Original implementation was used for recreation as much as feasible, along with extensions to build upon and verify the findings of the published research.

Introduction

Problem Statement

The ability for medical facilities to find early warning signs of medical emergencies is crucial for the well-being and advancement of society. FAMEWS, the abbreviation for a Fairness-Auditing tool for Early-Warning Systems, which researched the link between warning system accuracy in comparison to demographic discrepancy and was presented in the Conference of Health, Inference, and Learning (CHIL) has sought to identify where the line is drawn between true accuracy and error based on factors not taken into account (Hoche et al. 2024). Advancements in artificial intelligence have raised concerns about the bias embedded within and its validity in predicting outcomes in varying emergency medical scenarios. This project's reproduction aims to evaluate the credibility of the results of FAMEWS in regards to whether demographic consideration is effective in diagnostics. Along with this, there is a goal in expanding upon the current state with methods of a new dataset and feature removal to compute the magnitude of its effects in final results.

Citation to Original Paper

The original paper used for reproduction in this project is ”FAMEWS: a Fairness Auditing tool for Medical Early-Warning Systems” (Hoche et al. 2024). The full citation can be found in the references section at the end of the report.

Scope of Reproducibility

We were able to reproduce the core model testing that was evaluated in the study. A public healthcare dataset with simulated characteristics was used due to a verification process required to utilize the original HiRiD (High-time Resolution ICU Dataset) data that was incorporated in the full FAMEWS research (Yèche et al. 2022). The differentiating dataset resulted in conflicting in comparison to the hypothesis which is explained in later sections. Model training and fairness evaluation was done as close to the original study as feasible with the resources available to the public.

Methodology

Environment

The programming language used for reproduction was Python 3. LightGBM was the model used to train the gradient boosting classification models for outcome prediction. Frameworks and libraries used were scikit-learn for data preprocessing and model evaluation, Fairlearn for fairness metric computation, Pandas for dataset handling and manipulation, and Numpy for numerical operations. All testing was done on Google Colab. The language and libraries align with what was used in the FAMEWS publication. Overall the environment consisted of:

- Python
- LightGBM
- Scikit-learn
- Fairlearn
- Pandas
- NumPy

Data

We used a public healthcare dataset involving statistics of diabetes patients and converted it into binary classification. This dataset consisted of approximately 350 patient records with ten numerical features per patient. Features included, body mass index, blood pressure, insulin measurements, glucose level, along with other indicators that are linked to diabetes risk. Synthetic demographics were added including sex, age, and race to evaluate the original paper hypothesis that these characteristics impact results. The target variable in the dataset contained multiple outcome categories in relation to patient condition and was transferred to binary classification to signify adverse medical outcome vs average condition. The data split used was 80% training and 20% testing grouping to ensure equal class representation across both partitions. The repository for FAMEWS is public and available on GitHub as a repository at: `https://github.com/ratschlab/famews`. In the original research, the experimentation was done by utilizing data on the HiRID (High-time Resolution ICU Dataset) and MIMIC-III (Medical Information Mart for Intensive Care) and datasets, which provided statistics on patients stays including data such as their vital signs, lab test results, diagnoses, outcome, along with their demographics. However, due to the alternative dataset not containing demographic attributes necessary for testing, synthetic features were implemented. Metrics used in FAMEWS included AUC gap, equalized odds, and demographic parity difference. The same model and metrics have been used for reproduction. The synthetic characteristics and libraries involved allowed for efficient computation of these metrics while keeping protected patient record data private. While equivalent clinical data was not implemented in replication, the public dataset and implementations allowed for core testing equivalent to the methods used in the original research to evaluate the hypothesis that demographic features should be included in critical patient testing in early warning systems for the most feasibly accurate prediction.

Model

The model used was a LightGBM binary classifier. It is a gradient boosting framework constructed for efficient learning for structured data. This aligns well with the task of clinical patient prediction involving numerical attributes. Input involved health features of patients from a public diabetes dataset and output was the probability of adverse medical outcomes. This is the same model and method in the original research of FAMEWS, making a LightGBM model ideal for reproduction. Models were trained from scratch from the public health dataset.

Training

Hyperparameters

Hyperparameters for the LightGBM model were manually specified and remained the same across all experiments to ensure fair comaprison. This included the learning rate set to 0.05, 31 leaves, and 300 boosting rounds, along with the max depth being unrestricted. These values were selected so that replication could remain consistent between baseline, demographic, and extension training, and reduce tuning that could lead to bias within the results. Three training runs were conducted, one for the baseline model, one for demographics included, and one with the ablation of the age characteristic. Loss was optimized through gradient boost paired with cross entropy as the objective function, which aligned well with the task of patient outcome prediction. The decided setup ensured that results were impacted solely by demographic features, rather than hyperparameter tuning which could have further impacted results.

Computational Requirements

The replication was conducted using Google Colab and utilized only the CPU. The research was done with computation consisting of Pandas, Scikit-Learn, Fairlearn, and PyHealth. Fairness was audited in regards to a LGBM (Light Gradient Boosting Machine) early-warning systems for the alternative diabetes dataset. The LightGBM training runtimes consisted of a few seconds per run, with the total runtime for baseline, full demographic inclusion, and ablation experimentation remaining below one minute. Memory usage remains below the limits of Google Colab due to the small dataset size and the absence of deep neural network training. There was no GPU usage required since gradient boosting on tabular data is efficient utilizing a CPU. Each model was trained for 300 boosting rounds. The original research pipeline depended on large scale ICU data with extensive preprocessing and model evaluation, which would require much longer runtimes and hardware optimization for training. A smaller scale replication with the public diabetes dataset has allowed for core training replication which is most efficient for public reproduction without the original dataset. While this causes limits to full comparison to the original results, it enables a practical reproduction utilizing public hardware and data constraints. Overall, replication was conducted prioritizing core training while also being manageable to recreate through alternative public and simulation data.

Training Details

Training was done using binary classification and optimized through gradient boosting. There was one training run used for baseline configuration to sign all features available in the public diabetes dataset. Another training was conducted with three demographics of age, race, and sex, and one training run for the ablation of the age feature. This controlled setup ensured there was minimal risk of bias affecting the results. Due to the smaller dataset size compared to the original research, training was rather quick and did not require many adjustment techniques or subset usage. Loss was measured through binary classification loss. Overall, the objective of replicating the effect of demographics on patient outcomes was done through measuring predictive performance and fairness metric rather than extensive hyperparameter adjusting. This method allowed for efficient reproduction of the core study of FAMEWS, and assisted in proving that demographic features are capable of making an impact in accurate prediction of patient outcomes.

Evaluation

Model performance evaluation was conducted using Area Under the Receiver Operating Characteristic Curve (AUC). Fairness was evaluated using both a demographic parity difference and equalized odds difference. All of these metrics are consistent with the core performance and fairness measures of the original study. Training and testing was done on a 70% to 30% split for baseline, demographic, and ablation testing. These fairness metrics consisting of demographic parity difference and equalized odds difference allowed for analysis of fairness to accuracy tradeoff that is caused by demographic features. Analysis of prediction behavior indicates that the baseline model placed patients with extreme glucose and insulin values as strong prediction indicators.

However, considering this is a diabetes dataset this association is expected. Once ablation of age occurred, prediction confidence became closer across different samples. However age ablation did not reduce predictive performance as shown by the slightly higher AUC. This could suggest that correlated clinical features may impact results for removed demographic features. The evaluation of replication supports the core findings of FAMEWS about how demographic features can impact a model's predictions and fairness behavior in medical settings.

Results

Table/Figures

The next page displays Table 1: Model Performance and Fairness Metrics which shows results of AUC, demographic parity, and equalized odds at baseline, demographic inclusion, and ablation of age. The baseline model resulted in the lowest AUC which signals a slightly lower prediction performance. However the ablation results did not show a decline in accuracy after the removal of the age feature, likely due to the limited dataset. Demographic parity decreased with demographics and slightly increased after the ablation, while equalized odds remained the same. The table portrays the results of how demographic features impact results in clinical patient outcome prediction.

Compare and Contrast

When removing age in training, model accuracy was increased, demographic parity improved and equalized odds changed slightly. The observed increase in AUC after removing the age characteristic suggests it may have introduced noise with other clincal features considering they were implemented manually. Once removed, the model relies more on clinical features which causes a slight gain in predictive performance. The equalized odds also shifted slightly, which could be due to an indirect correlation between age and other clinical features being reduced. Due to this, it is expected that improvements in particular fairness metrics could result in a shift of other metrics, and should be accounted for before deployment to clinical settings. When comparing these results to the original research, the fairness gaps were smaller and less variable across the features of age and sex, likely due to the complex patient history associated with the dataset. The alternative diabetes dataset is much more static leading to a limiting consideration of large demographic disparities, though still maintains the fairness and accuracy results. Overall these results show that implementing demographic features in model testing does not always guarantee higher accuracy but demonstrates an accuracy to fairness tradeoff, which is consistent with the findings of the original research.

Extensions Ablations

One method of improving the FAMEW creators' methods includes utilizing the current state on a new dataset. This is acknowledged by researchers in the original report in Section 4.3 regarding strengths and limitations of the model. This was done by implementing an alternative dataset as mentioned previously, and resulted in testing of the range of the pipeline's functionality across a different dataset setting. This was done by utilizing the alternative diabetes dataset rather than the HiRiD dataset which was used in the original study. Along with this, a method of data visualization was implemented in the table section of this report where readers can quickly compare and contrast reproduction results in varying instances. Finally, removing an individual demographic feature such as age through ablation was used to see differentiating effects, allowing researchers to understand primary methods for accuracy and fairness in the pipeline. Generating subsets of data in the simulation dataset would allow an environment to observe how metrics such as equalized odds or demographic parity impact final evaluations. Further experimentation could reveal which fairness indicators are most impactful and would provide more security to real patients. Further variable adjustments would either improve the FAMEWS state through additional adjustments, or solidify the authenticity of the current auditing model. The ablation used for research was the removal of the age characteristic while retaining sex and race features. This tested the impact of age on model behavior. This ablation showed that removing age resulted in a slight increase in accuracy while fairness metrics changed slightly, showing that demographic features can influence fairness and accuracy in early warning systems. Overall, the methods implemented have the potential to assist in expanding the understanding of the limitations of FAMEWS across various members involved in its usage including researchers and medical professionals.

Discussion

Implications

Models can be accurate in testing, but unfair if not all characteristics are considered. The original FAMEWS study is only partially reproducible for public research due to restricted access to the HiRiD dataset. Core trends can be reproduced using public data and adjustments, but the alignment of the exact results appear unfeasible. Demographic information on patients is crucial for the most accurate evaluation of the likely outcomes of the patient, which is consistent with what is proposed in the original research. However, our results also show that removing demographic features can shift both accuracy and fairness, emphasizing the importance of careful evaluation before these results are considered in clincal setting. The novelty of this research involves the specification of the purpose of FAMEWS, in regards to the fact it is designed for the purpose of early-warning systems in medical environments and to be an extension of current fairness methods such as Fairlearn while including clinical attributes beyond the standard. Fairness dimensions included involve patient events, clinical outcomes, and sources of bias. This research is extremely relevant to current times, where a desire for accurate artificial intelligence in medical settings is in high demand for increased speed and accuracy in patient diagnoses and treatment. FAMEWS focuses on ensuring predictive models can be trusted for a variety of medical scenarios and provide the most accurate results. It was hypothesized that the state of the art early warning systems did

Table 1: Model Performance and Fairness Metrics

Model	AUC	Demographic Parity	Equalized Odds
Baseline	0.7967	0.0378	0.0297
Demographics Included	0.8189	0.0043	0.0847
Age Ablation	0.8272	0.0111	0.0847

not factor in characteristics such as age, sex, and type of admission, leading to possibly inaccurate findings in critical event discoveries using current early-warning systems. The belief was that integrating metrics and evaluation centered around unique medical scenarios could provide a clearer understanding of an individual patient's situation and more accurate diagnosis results. While the approach of utilizing alternative and synthetic data allows the opportunity of public replication, it does not allow exact and full replication of the original study. This is due to the lack of complex medical history which is provided in the HiRiD datasets and used for testing in FAMEWS. This results in metrics such as fairness gaps being smaller and less variable than the original findings. However the observed tradeoff between accuracy and fairness still follows the core trend reported within the original research. This suggests that even when smaller clinical datasets are utilized, demographic features still play a role in patient outcomes during model evaluation.

Ease of Reproduction

The replication process for model training was relatively accessible considering adjustments were made to conduct training through the use of an alternative public dataset. All experiments were conducted on Google Colab on CPU optimization without requiring accelerated GPU. The LightGBM training required a few seconds per run, and metric evaluation was also relatively quick. Due to the use of open source frameworks and libraries such as scikit-learn and Fairlearn, implementation was manageable at this scale.The ease of reproduction is primarily caused by the smaller dataset used, and considering if the original HiRiD was implemented for replication the hardware requirements and runtime would likely have been much longer and not practical for public replication. The alternative method allowed for core testing and analysis to be conducted while allowed for low hardware and runtime requirements.

Difficulty of Reproduction

The original HiRiD dataset is inaccessible to the public without verification, so the full clinical preprocessing pipeline could not be reproduced. This also implies that the full pipeline that is used on the dataset can not be utilized without access to the complete clinical data. The experimentation consisted of two datasets of HiRID and MIMIC-III for medical records of ICU patients. However, because these are datasets of real patients, they regulated and require credential verification for approved researchers, which could take many weeks or beyond. Along with this, the data used cannot be shared or posted anywhere without approval, so if using the original dataset for replication, it would not be approved to appear in final submission. This leads to the alternative method of using a public health dataset with simulated demographics, which was used in reproduction but could have impacted results. Along with this, the original research evaluated fairness across multiple random seeds and implemented SHAP based group attribution for a portion for fairness interpretation. Since the alternative dataset is limited, SHAP based group feature attribution was not implemented, as it was a minor portion for the study and too extensive for reproduction at this scale without a compatible dataset. While this did prevent the use of all metrics trained to the same depth of the original work, the core alignment of fairness and performance trends were still replicated. Overall due to these implications, while an exact replication is unattainable to the public, an alternative dataset still allowed meaningful core model replication and result analysis.

Recommendations

A potential recommendation could be that the authors should provide public demo datasets that can be used for testing and closely resemble the properties of the HiRiD data without revealing personal patient information. This would allow the public a clearer understanding of how training and evaluation is conducted and would allow for deeper analysis of results in relation to input. Providing simplified example pipelines and baseline experiments for users not in the medical field could allow for further research and replication to be done by a broader range of researchers.

Author Contributions

This project was completed by Fahad Zia.

References

Hoche, M.; Mineeva, O.; Burger, M.; Blasimme, A.; and Ratsch, G. 2024. FAMEWS: a Fairness Auditing tool for Medical Early-Warning Systems. In Pollard, T.; Choi, E.; Singhal, P.; Hughes, M.; Sizikova, E.; Mortazavi, B.; Chen, I.; Wang, F.; Sarker, T.; McDermott, M.; and Ghassemi, M., eds., *Proceedings of the fifth Conference on Health, Inference, and Learning*, volume 248 of *Proceedings of Machine Learning Research*, 297–311. PMLR.

Yèche, H.; Kuznetsova, R.; Zimmermann, M.; Hüser, M.; Lyu, X.; Faltys, M.; and Rätsch, G. 2022. HiRID-ICU-Benchmark – A Comprehensive Machine Learning Benchmark on High-resolution ICU Data. arXiv:2111.08536.