Cianna Grama
Professor Moher
DAT 201: Data Thematic Inquiry
17 April 2025

Utilizing Synthetic Testing Data to Enhance Software Reliability in the Medical Field

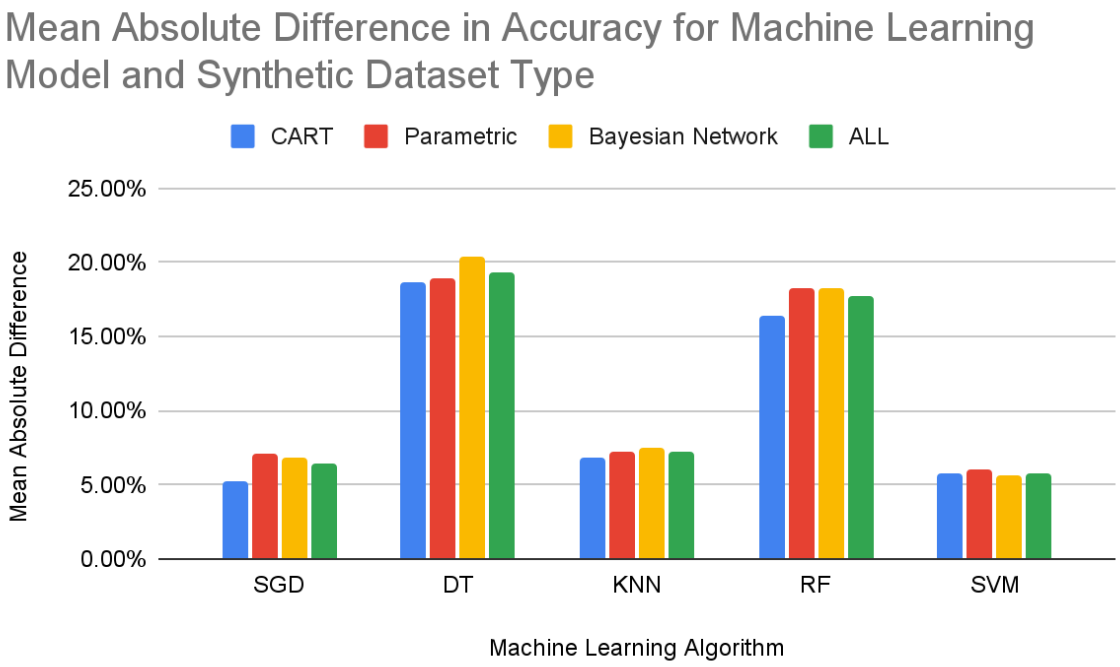
After artificial intelligence (AI)'s introduction to the world, it has been brought to many industries, including the medical field. A study in 2023 reports that 65% of US hospitals used predictive models with 79% using models from their electronic health records (EHR) (Nong et al.). As the use of predictive models increases in the medical field, it is increasingly important for these models to be accurately trained. Historically, predictive models are trained on real patient data, but there are several potential problems with this. One is that using patient data puts patient privacy at risk and potentially allows for the exploitation of the patient's data. Additionally, real patient data is often under-representative of minority populations, leading to under-representation in the training datasets of the predictive models. One example of this is the underrepresentation of black men in cancer clinical trials (Unger, et. al 2020).

A potential solution to these problems is using synthetic data to train these predictive models. Synthetic data, according to GenRocket, is entirely new data produced by controlled synthetic representations of what a data production environment would look like. In the medical field, this would look like simulating patients with their own illness(es), demographics, experiences, and outcomes. Synthetic data is generated by creating simulated situations where models interact to create the data or by applying sampling techniques to real data. As the demand for accurate predictive models increases, researchers need to advance and implement synthetic data generation within the field of medical software development because of their potential to create more efficient and inclusive data models that protect patient privacy and address issues of unrepresentative data sets common in medical models.

A 2020 study entitled "Reliability of Supervised Machine Learning Using Synthetic Data in Health Care," aimed to test the accuracy of using synthetic data to train medical models using machine learning, while maintaining patient privacy (Rankin et al., 2020). This research created synthetic versions of 19 publicly available healthcare datasets using 3 techniques: classification

and regression trees (CART[1]), parametric methods[2], and bayesian networks[3]. Then, for each data set, they tested 5 different machine learning classification models[4] by training the model on real data, training the model on synthetic data, and testing both trained versions on real data. In comparing the models based on accuracy, precision, and recall, the results showed that the synthetic data models were less accurate than the real data models, but not by a large amount. The SVM, SGD, and KNN models had the lowest mean absolute difference in accuracy, as seen in Figure 1. The consistently small mean absolute difference in these models suggests that synthetic data, as it currently stands, maintain a statistically insignificant decrease in accuracy than real data. The outcomes of this research show that synthetic data is a promising solution to combat the challenges of privacy and representation in medical software development. As the synthetic data outcomes are less accurate than the real data, more development is needed in this area. Even, so the differences in accuracy are not so extreme that they are unmanageable, and as privacy issues arise, the need for protection is imperative.

Figure 1



[1] a decision-tree-based method that generates synthetic data by learning patterns from real data
[2] using mathematical models (like distributions) to estimate data characteristics
[3] using probability models to generate data by capturing the relationships between variables
[4] support vector machine (SVM), K-nearest neighbors (KNN), stochastic gradient descent (SGD), decision trees (DT), and random forest (RF)

In a study examining synthetic data generation and evaluation techniques, researchers aimed to improve medical machine learning models by addressing the demographic imbalance in datasets that historically leads to poor predictive performance in minority groups (Bae et. al., 2025). The researchers started with real datasets with imbalances. Then, they generated synthetic data for the minority groups using synthetic minority oversampling techniques (SMOTE[5]) and gradually added the synthetic data to the training model. After testing the models on performance, if the newly added data improves model performance it was kept, otherwise, it was removed. This process was repeated until the model performance was refined properly. This method improved the models by 4.01% - 7.79% across the 3 SMOTE variants. The results of this study indicate that synthetic data is extremely useful in making medical machine-learning models more accurate for minority groups, especially in areas where it is difficult to gain real data.

The results of these two studies show the promise of synthetic training data to support medical software development and the urgency of investing in creating better synthetic data. The findings of Bae et al. (2025) show that synthetic data is extremely useful in filling in the gaps of underrepresented populations in datasets. The findings of Rankin et al. (2020) show that synthetic data training outcomes are only slightly less accurate than real training outcomes and that this gap is manageable, especially when privacy is at risk. Together, these findings suggest that utilizing synthetic data already wields powerful results and that further development and research in this field could potentially close the gaps and increase the success of synthetic data usage. As machine learning and predictive models are increasingly used in the medical field, privacy and representation must be preserved.

One potential critique of synthetic data is that generative models may not be able to replicate or capture the complexity and unpredictability of real data. In this argument, the arguably inaccurate synthetic data could potentially lead to inaccurate predictions by the models trained on the synthetic data. I would argue that this limitation is not a reason to not use synthetic data, but a reason to refine, grow, and improve it. If time and resources are put into refining synthetic data generation, there is potential to close the gaps and limit the limitations. The

[5] SMOTE generates new data points for the minority class by creating new data between the existing minority samples

benefits of using synthetic data - privacy and correcting underrepresentation among them - are significant enough that research on improving synthetic data is worthwhile.

As the use of predictive models in the medical field increases, it is crucial to advance synthetic data generation models in order to maintain the privacy of real patients and cover the underrepresentation of minority groups in medical training data. Looking forward, it would be useful for researchers to examine how synthetic data can be used to improve research, predictions, and diagnoses of rare diseases since there are often limited data sets on these diseases. In conclusion, researchers should focus on advancing and applying synthetic data generation in medical software development to build more accurate, inclusive, and privacy-conscious predictive models that overcome privacy challenges and underrepresentation in real-world data.

References

Bae, Wan D., et al. “Synthetic Data Generation and evaluation techniques for classifiers in data starved medical applications.” IEEE Access, vol. 13, 2025, pp. 16584–16602, https://doi.org/10.1109/access.2025.3532222.

GenRocket. “Adopting a New Synthetic Data Paradigm for Software Testing.” GenRocket Blog, 3 Oct. 2024, www.genrocket.com/blog/adopting-a-new-synthetic-data-paradigm-for-software-testing/.

Nong, Paige, et al. “Current use and evaluation of artificial intelligence and predictive models in US hospitals.” Health Affairs, vol. 44, no. 1, 1 Jan. 2025, pp. 90–98, https://doi.org/10.1377/hlthaff.2024.00842.

Rankin, Debbie, et al. “Reliability of supervised machine learning using synthetic data in health care: Model to preserve privacy for Data Sharing.” JMIR Medical Informatics, vol. 8, no. 7, 20 July 2020, https://doi.org/10.2196/18910.

Unger, Joseph M, et al. “Representativeness of black patients in cancer clinical trials sponsored by the National Cancer Institute compared with Pharmaceutical Companies.” JNCI Cancer Spectrum, vol. 4, no. 4, 24 Apr. 2020, https://doi.org/10.1093/jncics/pkaa034.