1. Define your research scope
   1. Aim to use machine learning and possibly deep learning to identify possible fraudulent trends or activity using public Medicare Data.
   2. We plan to use 3 years of Medicare part b and part d data along with labels joined from the LEIE to generate some insights about Medicare FWA. Since there are few fraud cases in the vast amount of Medicare data, we plan on mining these cases for insight and then using unsupervised techniques to identify fwa behavior.
   3. The data sets are quite large and combining them can be a bit of an issue. The public Medicare data sets are not released at the patient/event level. They are instead aggregated on NPI and other characteristics [13], [others I’m sure]. This means our team will have to join some large datasets together to get access to the relevant features we need.
2. Previous Research
   1. Anomaly detection is employed in many different areas such as procurement fraud [17], credit card fraud [18], and Medicare fraud detection [11], [16]. The assumption being that anomalous events or activity is likely to be fraudulent when compared with the rest of the body [13]. Teams have used Spatial Density using imLOF (Improved Local Outlier Factor) [11]. As well as unsupervised methods such as Isolation Forest and Unsupervised Random Forest [2], Deviation Clustering and GMM [16], as well as Bayesian Co-clustering [15]. Further past teams have seen that Local Outlier Factor (non-improved), K-Nearest Neighbors, and autoencoders are suboptimal performers [2]. Though there is some debate over LOF. [11], [2] The researchers involved in the study "An Anomaly Detection Method for Medicare Fraud Detection" designed a new LOF metric designated imLOF for improved Local Outlier Factor. This metric is designed to detect excessive medical treatment and decomposing hospitalization using spatial density information. [11] The original measure, developed by Breunig et al. [citation from 11.2], gives anomaly scores based on the density of observations, they noted that the density of anomalous event would less than that of its normal neighbors. This metric has issues with small clusters. They give an example of a small cluster of hypertension patients with a great deal of fraud. Since the cluster is small and the point’s neighbors are also likely to be anomalous the metric scores a low chance of anomalous activity. The authors then suggest an improvement to the LOF score by adding in the size of a cluster into consideration instead of only density, with the additional use of the DBSCAN algorithm the improved LOF score performed much better on healthcare data [11]
   2. Past research has seemingly only focused on single years [2], [12], [13], [etc] and used either a Supervised Learning design [6] or a combination of unsupervised and supervised [2], [8], [13]. Our design will follow the latter. There have also been studies direct at specific portions of Medicare/Medicaid such as dental [16], otolaryngology [5], and dermatology [12].
3. R. Bauder and T. Khoshgoftaar, "A Survey of Medicare Data Processing and Integration for Fraud Detection," *2018 IEEE International Conference on Information Reuse and Integration (IRI)*, 2018, pp. 9-14, doi: 10.1109/IRI.2018.00010.
4. R. Bauder, R. da Rosa and T. Khoshgoftaar, "Identifying Medicare Provider Fraud with Unsupervised Machine Learning," *2018 IEEE International Conference on Information Reuse and Integration (IRI)*, 2018, pp. 285-292, doi: 10.1109/IRI.2018.00051.
5. Bauder, R. A., & Khoshgoftaar, T. M. (2018). The effects of varying class distribution on learner behavior for Medicare fraud detection with imbalanced big data. *Health Information Science and Systems*, *6*(1), 1–14. https://doi.org/10.1007/s13755-018-0051-3
6. Buppert, C. (2001). Avoiding Medicare Fraud part 2. *The Nurse Practitioner*, *26*(2), 34–41. https://doi.org/10.1097/00006205-200102000-00005
7. Gad, B., Warren, C., & Paskhover, B. (2020). Why Otolaryngologists Get Excluded From Medicare and Medicaid. *Ear, Nose, & Throat Journal*, 14556132093304–145561320933040. https://doi.org/10.1177/0145561320933040
8. J. Hancock and T. M. Khoshgoftaar, "Performance of CatBoost and XGBoost in Medicare Fraud Detection," *2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA)*, 2020, pp. 572-579, doi: 10.1109/ICMLA51294.2020.00095.
9. Meyers, T. J. (2017). Examining the network components of a Medicare fraud scheme: the Mirzoyan-Terdjanian organization. *Crime, Law, and Social Change*, *68*(1), 251–279. https://doi.org/10.1007/s10611-017-9689-z
10. Musal, R. M. (2010). Two models to investigate Medicare fraud within unsupervised databases. *Expert Systems with Applications*, *37*(12), 8628–8633. https://doi.org/10.1016/j.eswa.2010.06.095
11. Nicholas, L. H., Segal, J., Hanson, C., Zhang, K., & Eisenberg, M. D. (2019). Medicare beneficiaries' exposure to fraud and abuse perpetrators.*Health Affairs, 38*(5), 788-793,793A-793C. doi:http://dx.doi.org/10.1377/hlthaff.2018.05149
12. Pande, V., & Maas, W. (2013). Physician medicare fraud: Characteristics and consequences.*International Journal of Pharmaceutical and Healthcare Marketing, 7*(1), 8-33.doi:http://dx.doi.org/10.1108/17506121311315391
13. W. Zhang and X. He, "An Anomaly Detection Method for Medicare Fraud Detection," *2017 IEEE International Conference on Big Knowledge (ICBK)*, 2017, pp. 309-314, doi: 10.1109/ICBK.2017.47.
14. Machine Learning and the Future of Medicare Fraud Detection - Gordon and Siegel
15. Machine Learning Algorithms with Big Medicare Fraud Data - R Bauder
16. An interactive Machine Learning Based Electronic Fraud and Abuse Detection System in Healthcare Insurance - Kose, Gokturk, and Killic
17. Application of Bayesian Methods in Healthcare Fraud - Ekin, Leva, Ruggeri, and Soyer
18. Outlier detection in healthcare fraud: a Case study in the Medicaid dental domain - Capelleveen, Poel, Mueller, Thornton, and Van Hillgersberg
19. Westerski, Kanagasabai, R., Shaham, E., Narayanan, A., Wong, J., & Singh, M. (2021). Explainable anomaly detection for procurement fraud identification—lessons from practical deployments. International Transactions in Operational Research, 28(6), 3276–3302. <https://doi.org/10.1111/itor.12968>
20. Moschini, Houssou, R., Bovay, J., & Robert-Nicoud, S. (2020). Anomaly and Fraud Detection in Credit Card Transactions Using the ARIMA Model.
21. Markus M Breunig, Hans Peter Kriegel, Raymond T Ng, and Jrg Sander. Lof: identifying density-based local outliers in ACM SIGMOD International Conference on Management of Data May 16-18 2000 Dallas, Texas, Usa, pages 93-104, 2000.