IMPACT OF ANOMALY DETECTION ON MACHINE LEARNING MODELS IN THE PREDICTION OF COVID-19 USING CATEGORICAL DATA

Presented by

ARVIND R NAIR

LIVERPOOL JOHN MOORES UNIVERSITY

Under supervision of

AMULYA DASH

INTRODUCTION

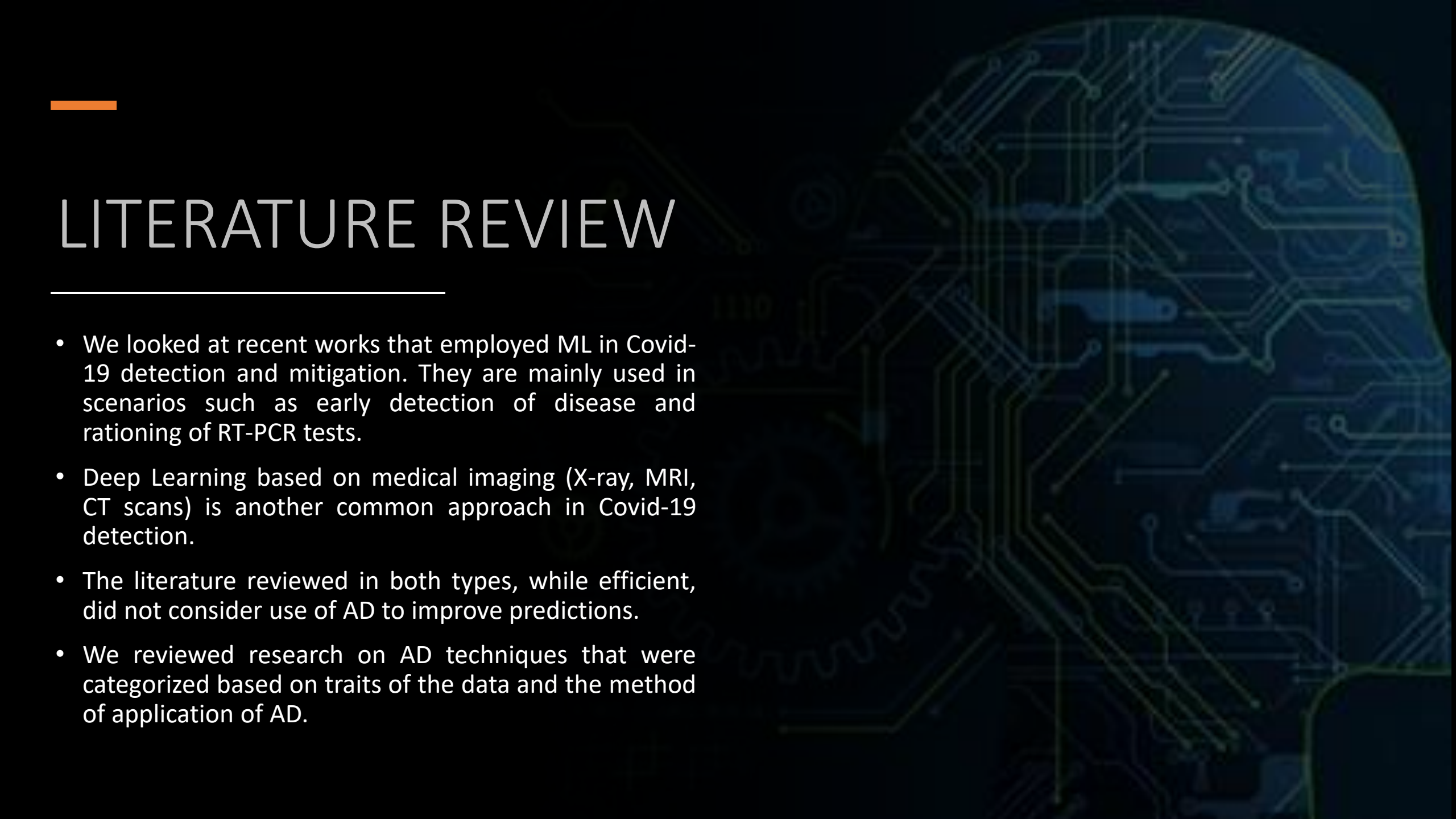
- The Covid-19 pandemic has resulted in wide-ranging impacts in public health, healthcare systems, economic health and in many other aspects of importance
- Early intervention through detection and mitigation of Covid-19 is a primary strategy for containment of the pandemic
- Our study found that research into ML for Covid-19 detection is increasing but use of AD to assist in the process is limited
- AD has innumerable applications in a variety of domains including healthcare and therefore can be considered for improving Covid-19 detection

INTRODUCTION

- Our research found that use of AD to assist in prediction of Covid-19 using a categorical dataset is yet to be explored
- The study proposes the application of Anomaly Detection (AD) as an integral part of training Machine Learning (ML) models to predict Covid-19 using a categorical dataset
- Use of AD to improve effectiveness of ML models has the potential to improve the strategy of early detection and mitigation of Covid-19
- Our study makes use of an anonymized categorical Covid-19 dataset which includes attributes on patient symptoms, gender, isolation treatment and travel history.

LITERATURE REVIEW

- We looked at recent works that employed ML in Covid-19 detection and mitigation. They are mainly used in scenarios such as early detection of disease and rationing of RT-PCR tests.
- Deep Learning based on medical imaging (X-ray, MRI, CT scans) is another common approach in Covid-19 detection.
- The literature reviewed in both types, while efficient, did not consider use of AD to improve predictions.
- We reviewed research on AD techniques that were categorized based on traits of the data and the method of application of AD.



LITERATURE REVIEW

- Developing AD solutions in categorical data was found to have received relatively less consideration
- Ensemble-based AD such as Isolation Forest were particularly focused on because of versatility in using them both for quantitative and categorical data
- Application of AD in Covid-19 was found to be limited. Research has yet to explore methods of improving predictive power when a structured Covid-19 categorical dataset is employed

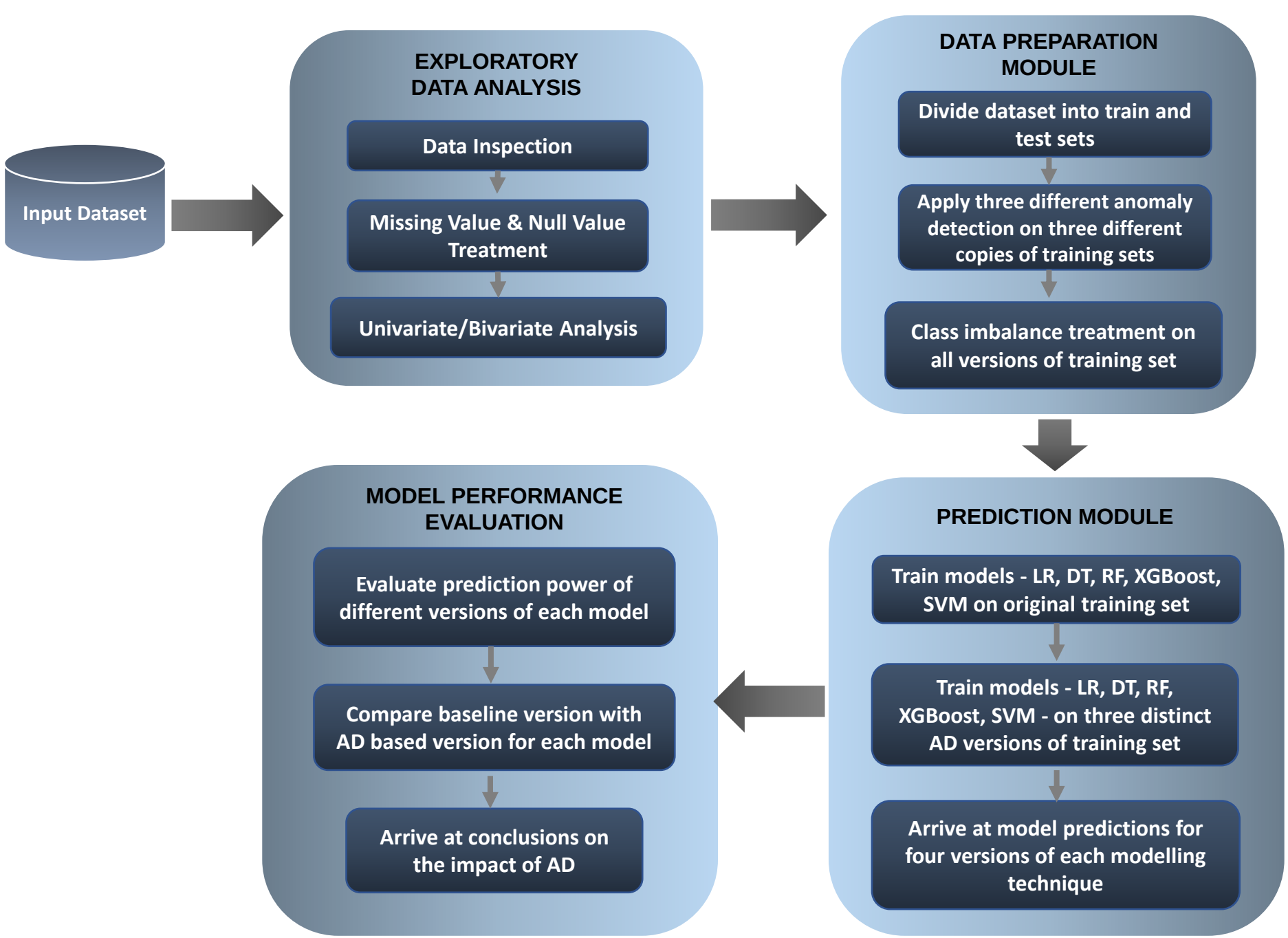
AIM & OBJECTIVES

The goal of the research is to investigate whether AD can improve the quality of data that is categorical in nature and thereby increase the effectiveness of ML models in predicting Covid-19 at an early stage.

The objectives of the research are stated as follows:

- To suggest an Anomaly Detection technique that is suitable for application on the Covid-19 dataset that is categorical in nature.
- To evaluate the impact of the Anomaly Detection technique on the predictive power of five different machine learning models, specifically Logistic Regression, Decision Tree, Random Forest, XGBoost and Support Vector Machines.
- To improve the prediction of Covid-19 at an early stage using the symptomatic data.

METHODOLOGY



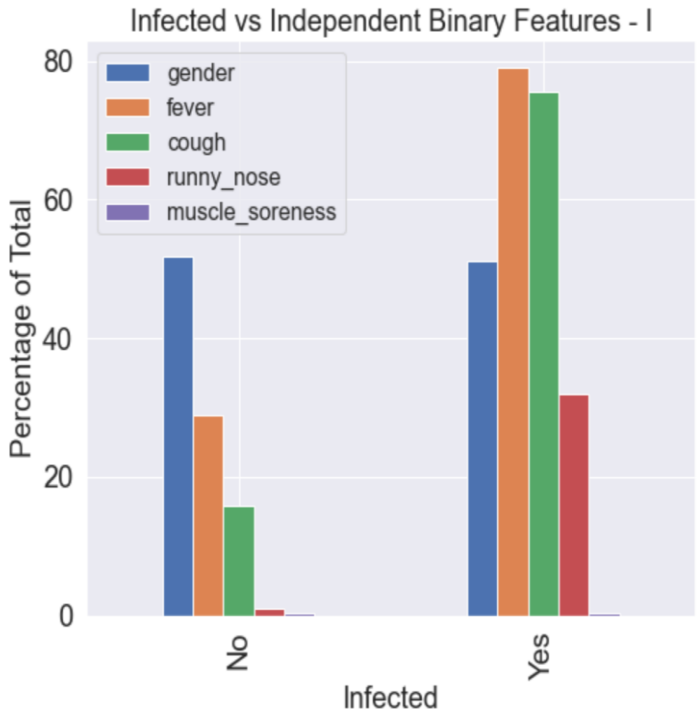
METHODOLOGY – AD TECHNIQUES

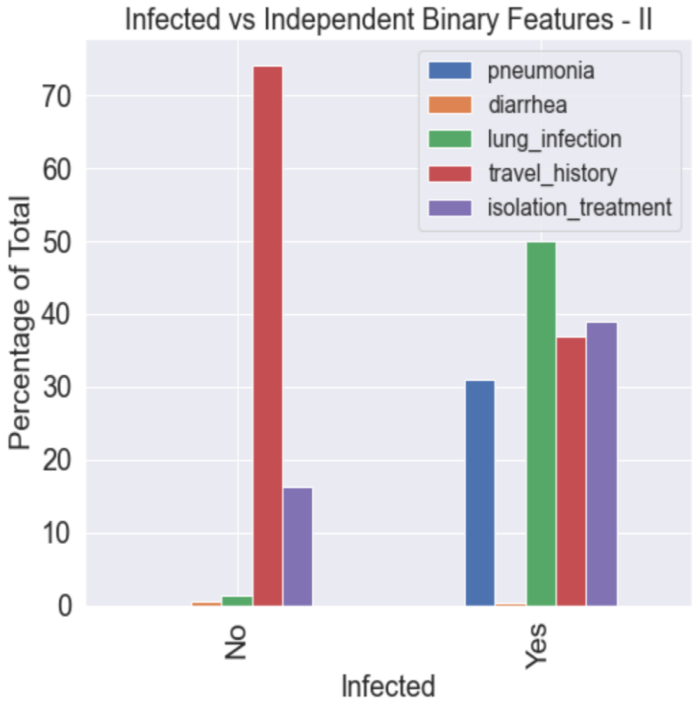
Three AD techniques were tested out in out methodology and are listed below:

- Isolation Forest (IF): An ensemble of trees based AD technique suitable for categorical/binary data. It returns anomaly scores and anomaly labels.
- Extended Isolation Forest (EIF): A special case of IF with improvements in detection of anomalies. It returns anomaly scores only.
- Histogram-based outlier detection (HBOS): An AD technique that bins data and creates histograms to identify anomalies. It returns scores and labels.

ANALYSIS & RESULTS – EDA

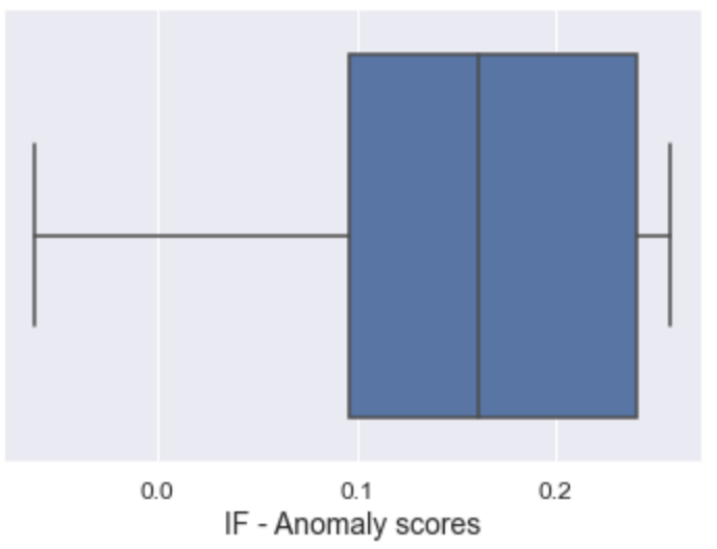
- We can see that Covid-19 infection does not affect one gender differently than the other
- Far higher patients (about 80%) with Covid-19 had fever and cough as symptoms
- Runny nose and pneumonia were also differentiators and indicated Covid-19 infection
- Travel history was not indicative of infection risk

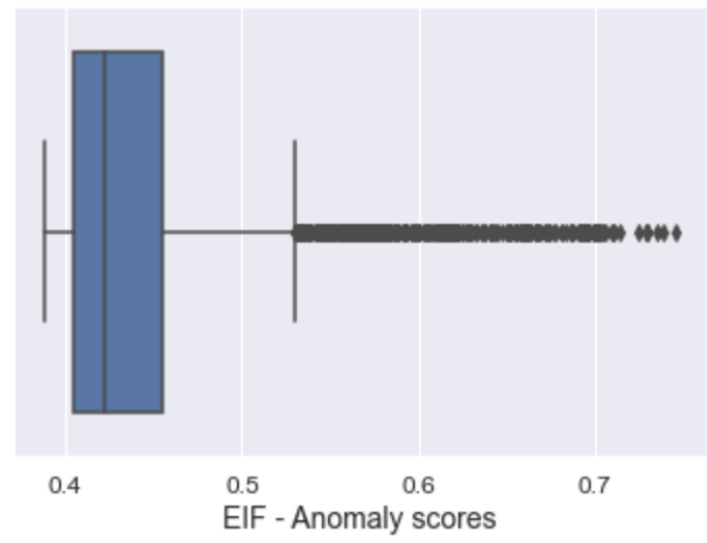


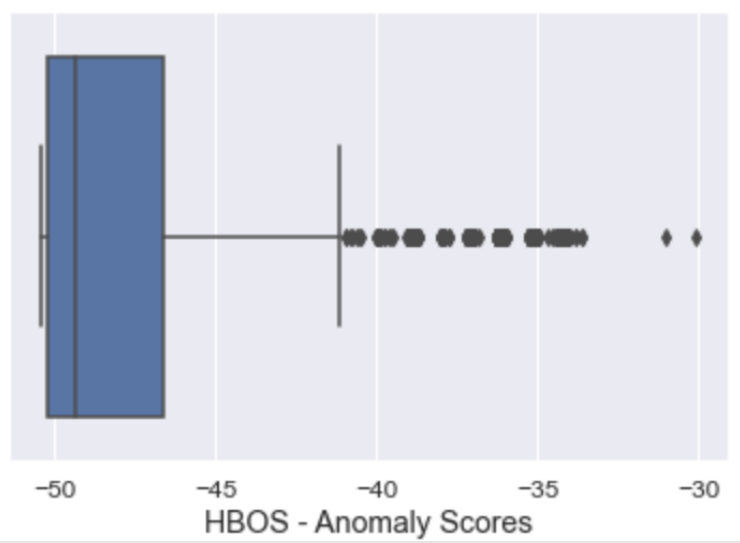


ANALYSIS & RESULTS - ANOMALY SCORES

- Datapoints with scores less than 0 were considered anomalies for IF while higher scores meant increased likelihood of being anomalies for EIF and HBOS.
- IF and HBOS automatically returned anomaly labels based on ‘contamination_factor’.
- Identification of anomalies using EIF required manually defining anomaly threshold.







ANALYSIS & RESULTS – TRAINING SET

- F1 score was used to tune performance of the models.
- AD with EIF performed best among three AD versions across all models.
- 3 of out 5 predictive models benefited from AD and LR benefited the most.
- DT and SVM saw decreases in performance after AD.

Comparison of Predictive Models – Training Set - Baseline Model vs Best AD based Model					
Classifier	**Without AD/ Best AD Method**	**Accuracy**	**Precision**	**Recall**	**F1 Score**
LR	Without AD	0.821	0.848	0.781	0.813
	AD With EIF	0.867	0.848	0.894	**0.870**
DT	Without AD	0.884	0.698	0.899	**0.786**
	AD With EIF	0.883	0.691	0.892	0.779
RF	Without AD	0.869	0.664	0.909	0.767
	AD With EIF	0.888	0.698	0.908	**0.789**
XGBoost	Without AD	0.888	0.706	0.900	0.791
	AD With EIF	0.894	0.720	0.883	**0.793**
SVM	Without AD	0.881	0.686	0.919	**0.785**
	AD With EIF	0.880	0.677	0.913	0.777

ANALYSIS AND RESULTS – TESTING SET

- Best performing AD versions differed depending on the modelling technique.
- AD with IF performed best of the three AD techniques.
- 4 out of 5 predictive models benefited from AD and LR once again benefited most.
- Only SVM saw decrease in performance after AD.

Comparison of Predictive Models – Testing Set - Baseline Model vs Best AD based Model					
Classifier	**Without AD/ Best AD Method**	**Accuracy**	**Precision**	**Recall**	**F1 Score**
LR	Without AD	0.833	0.639	0.770	0.699
	AD With IF	0.858	0.663	0.888	**0.759**
DT	Without AD	0.877	0.708	0.872	0.781
	AD With IF/HBOS	0.885	0.727	0.868	**0.791**
RF	Without AD	0.865	0.678	0.888	0.769
	AD With EIF	0.876	0.706	0.866	**0.778**
XGBoost	Without AD	0.877	0.707	0.874	0.782
	AD With IF	0.886	0.734	0.862	**0.793**
SVM	Without AD	0.861	0.673	0.874	**0.760**
	AD With HBOS	0.852	0.653	0.884	0.751

CONCLUSIONS

- Impact analysis of the differences in performance of five predictive models when AD is administered was completed as part of second objective
- Based on the testing set, we identified IF as an AD technique compatible with our categorical dataset while enhancing performance for four of our five models. This completed our first objective
- We partially fulfilled the third objective which was to improve the prediction of Covid-19. AD based predictions made improvements for all models except SVM
- However, outcomes of the research were restricted the limited size of dataset (6,512 samples). In future studies, a larger dataset is recommended

CONCLUSIONS

- Further research into other compatible AD techniques using larger datasets is required to find options for SVM.
- XGBOD is an ensemble-based AD technique that can be considered in future studies once documentation and implementation options expand.

THANK YOU