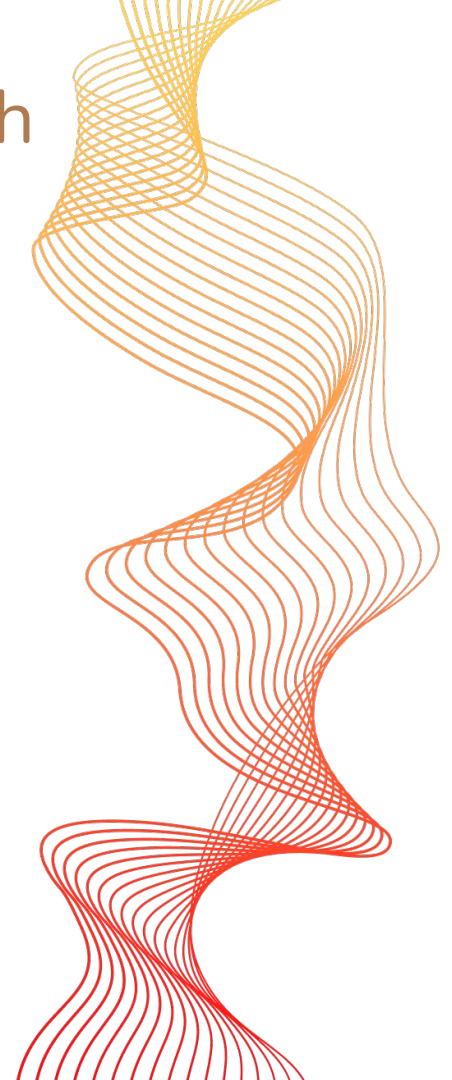


▶ Thyroid Disease prediction through Machine learning techniques with enhanced datasets

- Thyroid gland plays an important role in regulating metabolism, heart rate & body temperature
- A reduction in the production of relevant hormones (T3, T4) may lead to hyperthyroidism and hypothyroidism.
- The aim of this study is to analyse a popular Thyroid Dataset & increase the number of target classes used from 5 to 7 & 10 while retaining performance



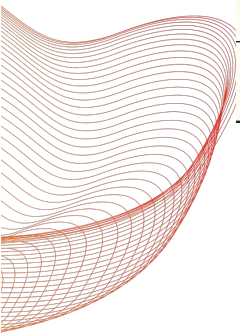


Existing work on Dataset

Existing studies on dataset have achieved 0.99 classification prediction score

ML models have outperformed DL models due to dataset size constraints

Best performing model is Random Forest Classifier with 200 trees and max depth of 20

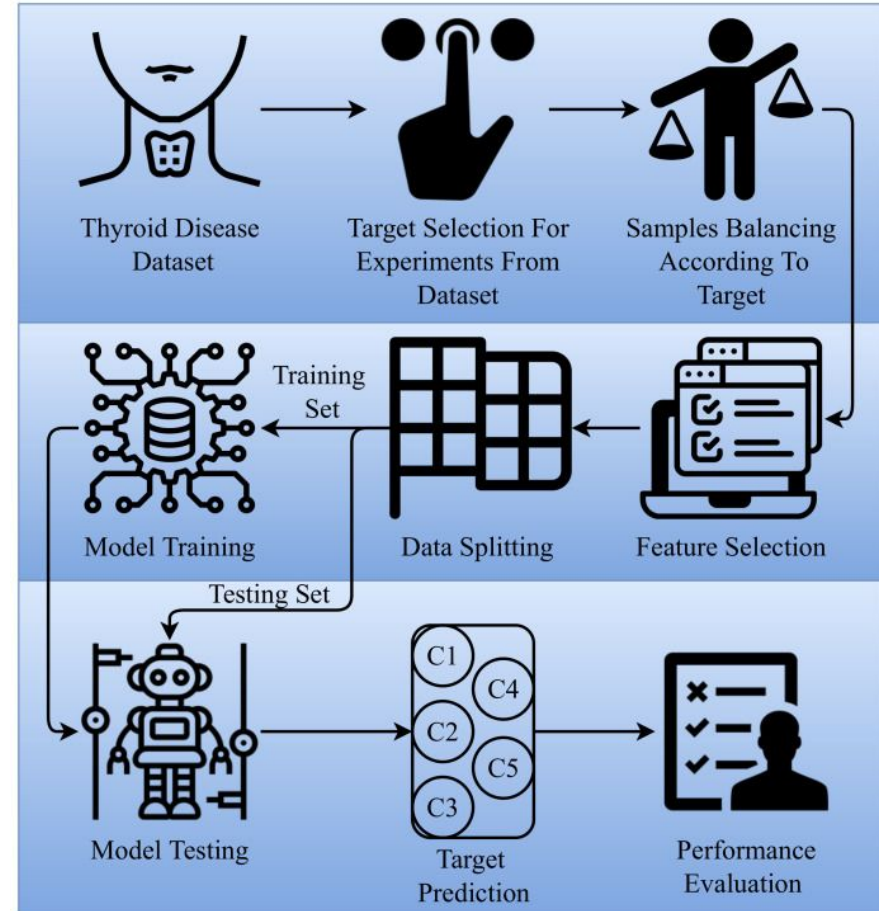


Class	Hyper-Parameters	Tuning Range
RF	n_estimators = 200, max_depth = 20	n_estimators = {10 to 300}, max_depth = {2 to 50}



Methodology for Project

- Dataset acquisition
- Preprocessing (sampling, encoding)
- Feature selection
- Model training & testing
- 10 fold cross-validation
- F1 scores
- Visualizations
- Writing paper



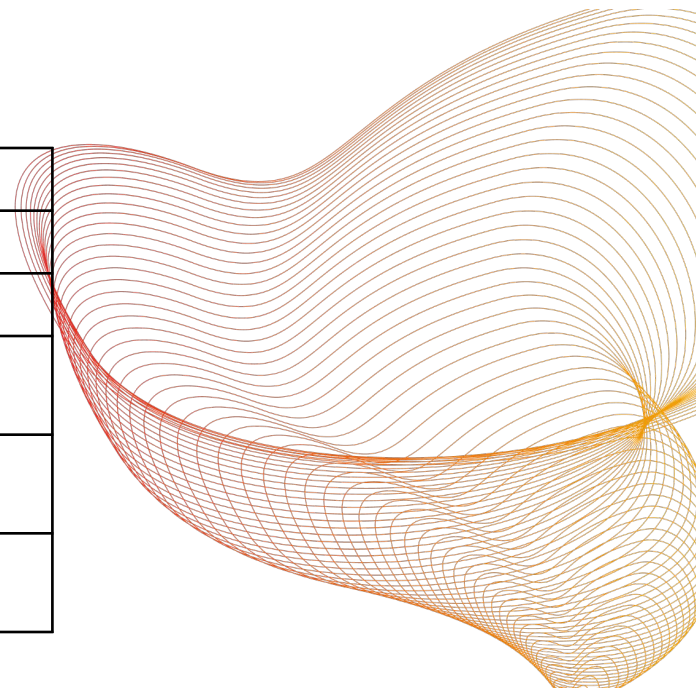


Dataset & EDA

Number of Samples	Number of features (bool /num)	Number of target labels
9172	30 (19 / 11)	29

- 5 target classes selected for training
- Class balancing done to avoid bias in training model

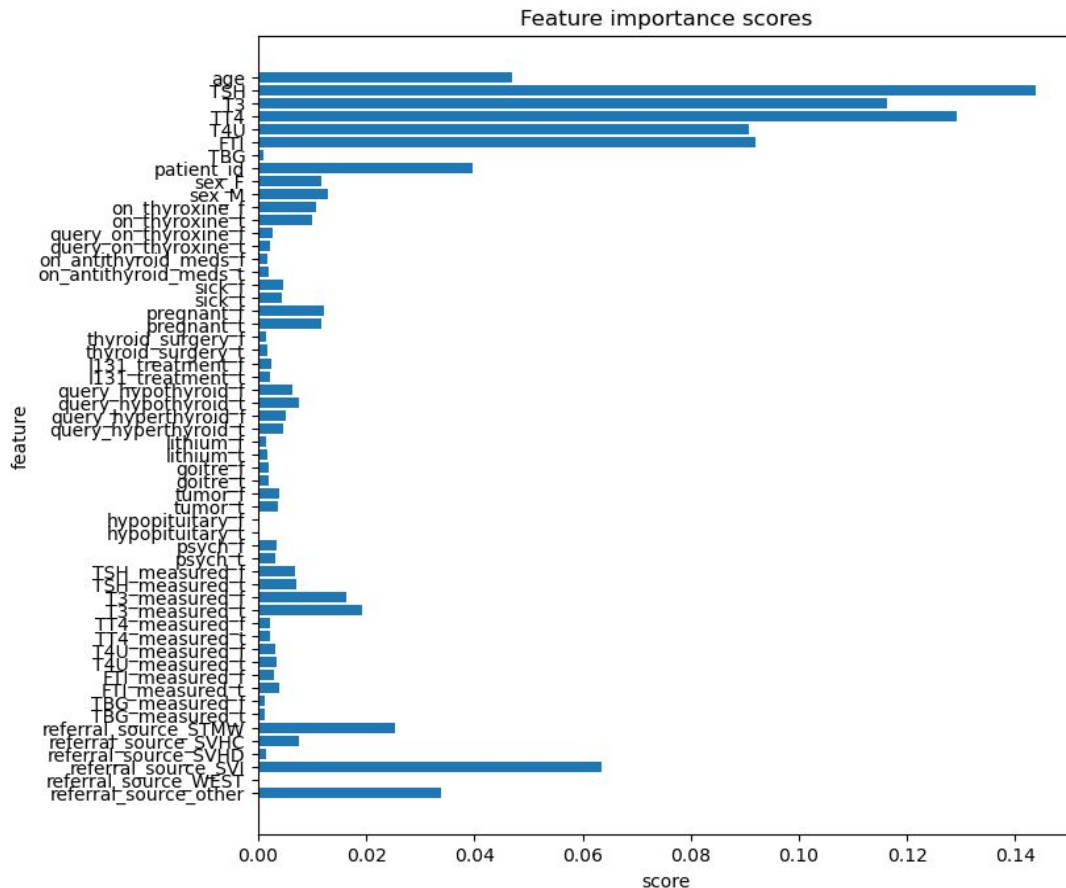
Class Name	Target Label	Original Count	Final Count
Normal	-	6771	400
Primary Hypothyroid	F	233	233
Increased binding protein	I	346	346
Compensated Hypothyroid	G	359	359
Concurrent non-thyroidal illness	K	436	436





Feature selection & model training

- Feature importance scores calculated
- Threshold value of 0.015
- 11 features used in all models
- Dataset split and trained on RFC algorithm

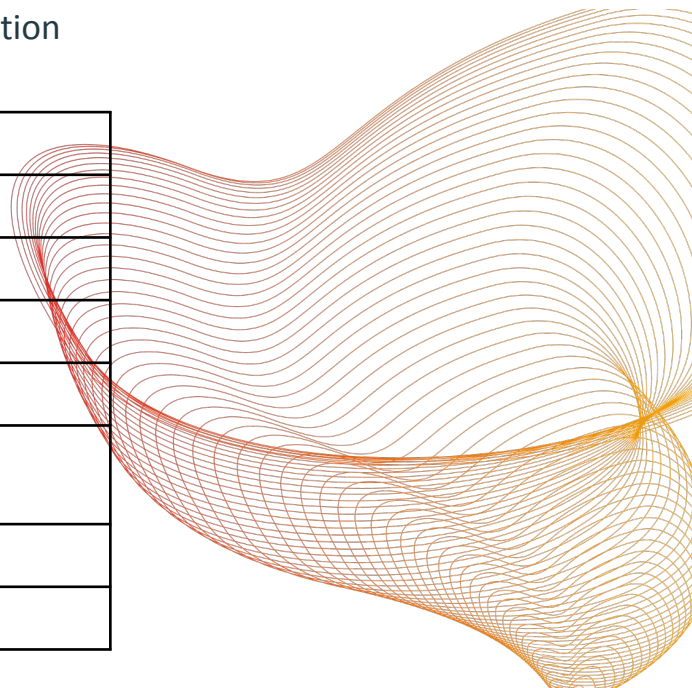




Synthetic Minority Over-sampling Technique (SMOTE)

- Data sampling technique which works by interpolating between existing data points and synthesising new samples
- Increased target classes from 5 to 7
- Trained on same parameters RFC model for consistency in evaluation
- Dataset features -

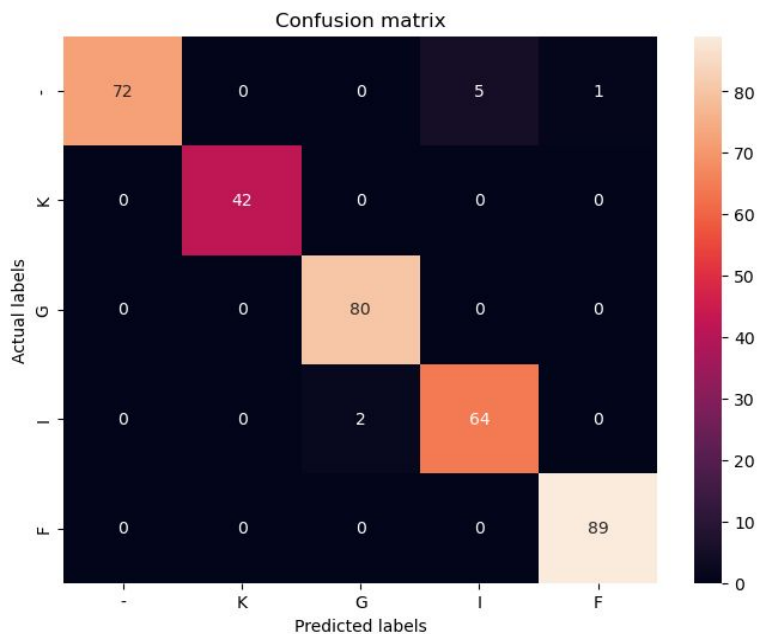
Class Name	Target Label	Count
Normal	-	400
Primary Hypothyroid	F	233
Increased binding protein	I	346
Compensated Hypothyroid	G	359
Concurrent non-thyroidal illness	K	436
Discordant assay results	R	231
Hyperthyroid	A	231



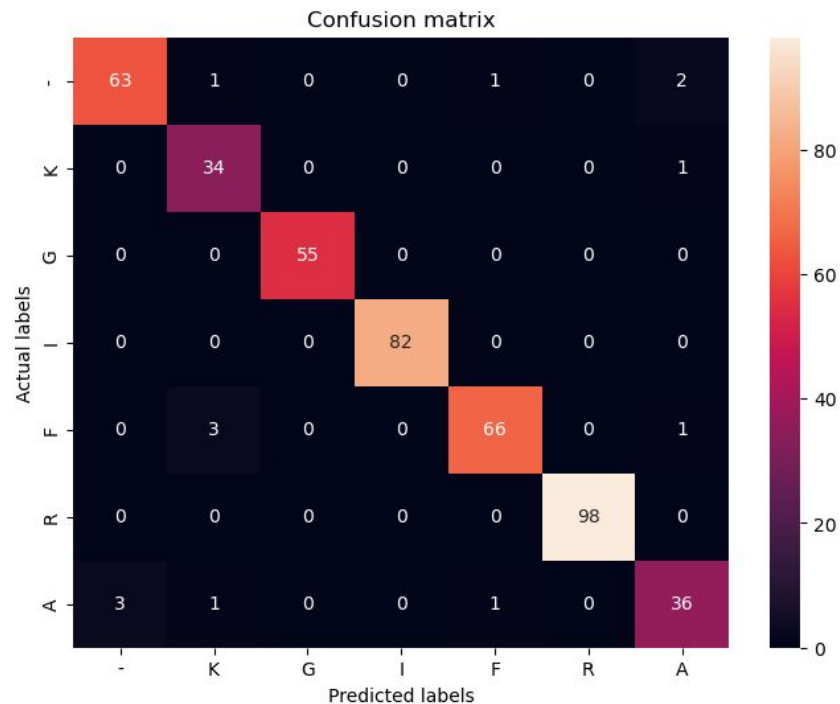


Results - Original vs SMOTE

Original model - **0.9786** prediction performance



SMOTE sampled model - **0.9629** prediction performance



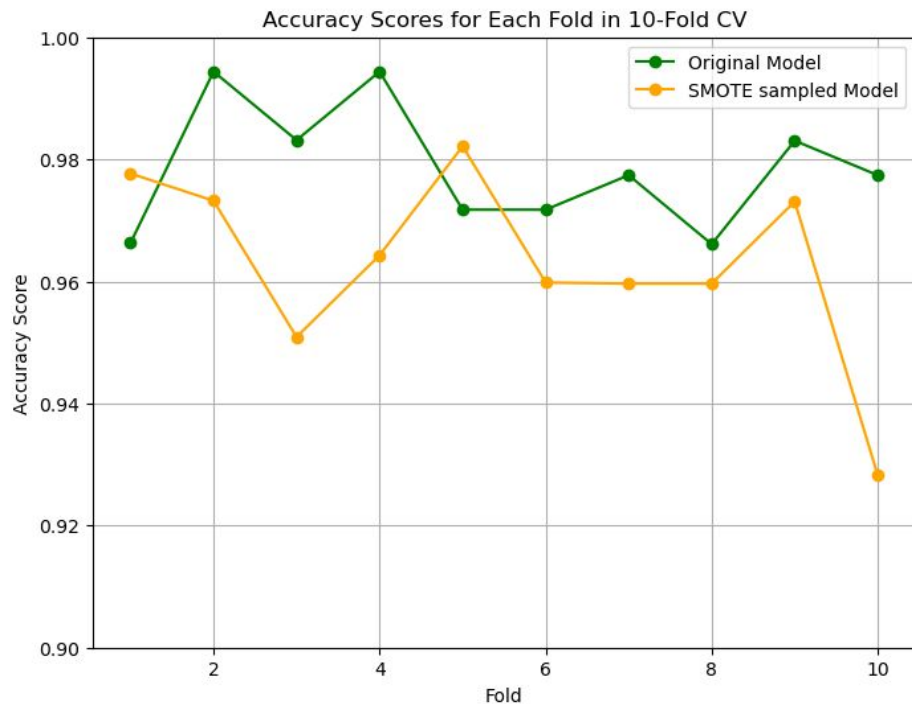


Results (contd.)

Original model

Mean accuracy - 0.9786

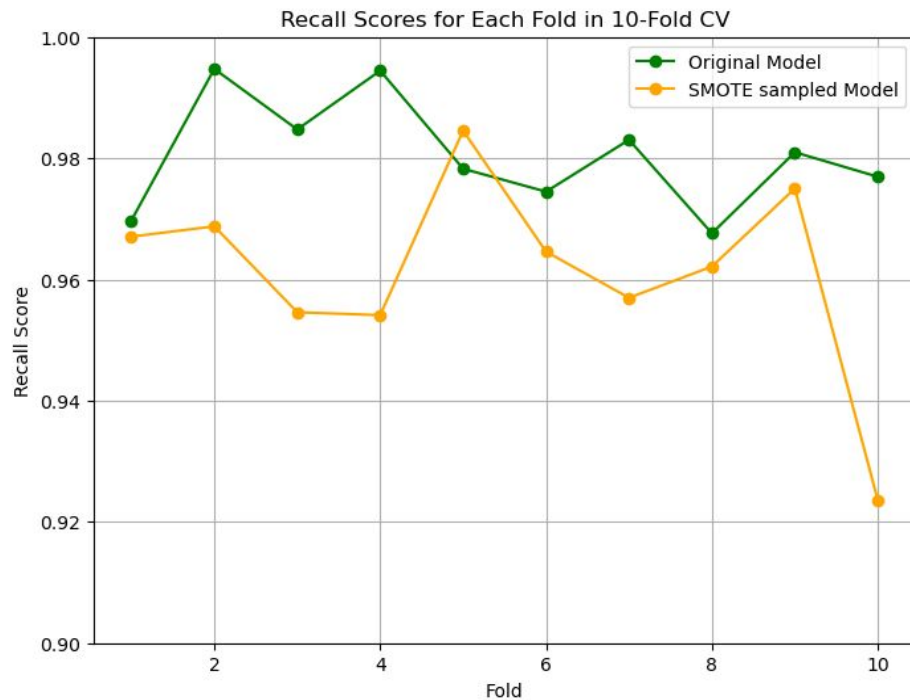
Mean Recall - 0.9805



SMOTE sampled model

Mean accuracy - 0.9629

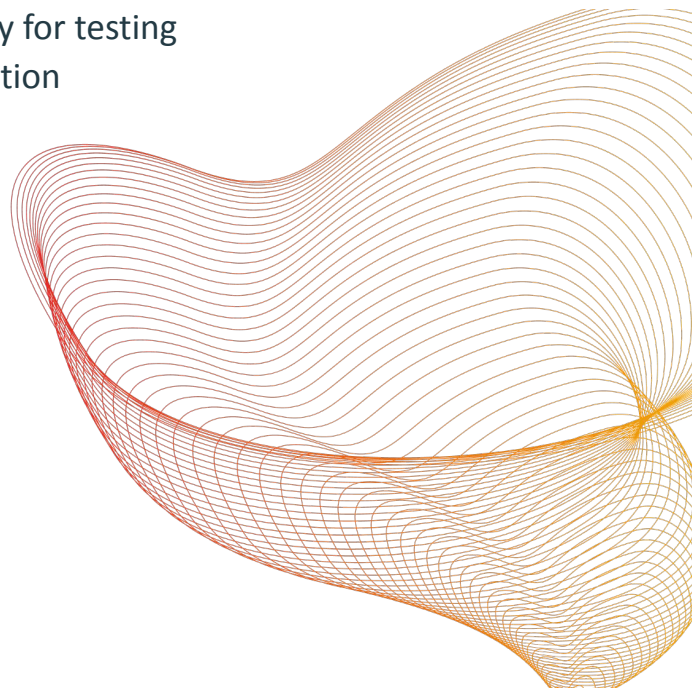
Mean recall - 0.9611





Borderline SMOTE Technique

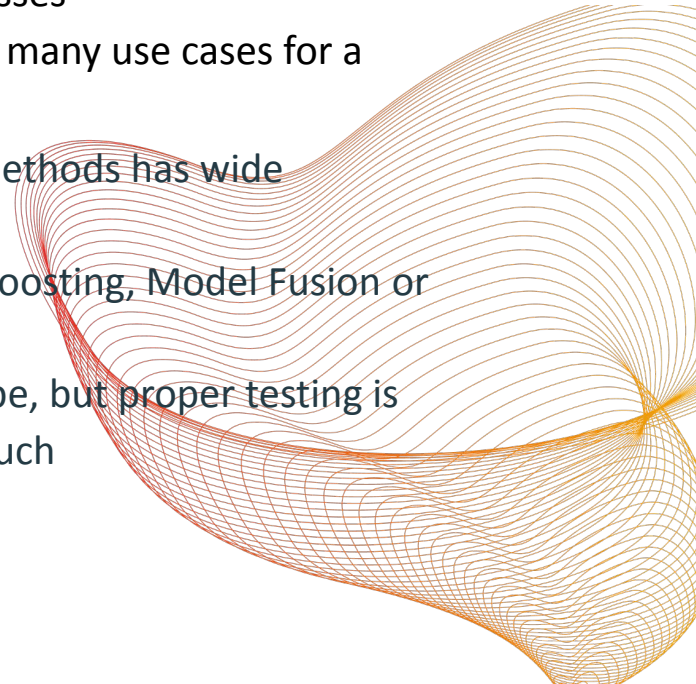
- Data sampling strategy used in augmenting datasets by synthesising new samples proximal to the decision boundary
- Increased target classes from 5 to 10, with approx 6500 samples each
- Removed original samples from training set to be used exclusively for testing
- Trained on same parameters RFC model for consistency in evaluation
- Prediction score of 0.8807 while identifying 10 classes





Future Research Scope & Conclusion

- Desired results from each ML Model are different
- One model may prioritise best performance scores, another may prioritise the ability to classify a larger set of data points or unique classes
- Nature of medical datasets - high in specificity, there are many use cases for a broader categorisation of classes
- Enhancing dataset richness through sampling or other methods has wide application in medical fields
- Future research can employ methods such as stacking, boosting, Model Fusion or transfer learning to enhance models
- Sampling methods can improve model classification scope, but proper testing is required to ensure model performance doesn't suffer much





Thank you
Aryan Malhotra - 2019ABPS0893P