The common cold, a viral infection affecting the throat and nose, alters speech characteristics due to inflammation and irritation of the vocal cord (Park, 2021). This study aims to develop machine learning models to accurately predict the presence of a cold based on speech patterns. The dataset utilised for this task is sourced from the INTERSPEECH Computational Paralinguistics Challenge Series. It comprises a total of 28,652 speech samples partitioned into training, development, and test sets for the task of classifying the presence or absence of cold based on 88 features extracted from each sample using the extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS). The training set consists of 9,505 samples (970 cold, 8,535 not cold), the development set includes 9,596 samples (1,011 cold, 8,585 not cold), and the test set contains 9,551 samples (895 cold, 8,656 not cold).

If one class significantly outweighs the other, conventional metrics like accuracy can be biased as a high accuracy score could be achieved by simply predicting the majority class for all instances. Unweighted Average Recall (UAR) ensures that the performance of the model is not skewed by the majority class, i.e., 'Not Cold' /negative class which outweighs the 'Cold' or positive class. UAR calculates the recall for each class individually and then averages them without weighting according to their respective class frequencies, thereby preventing bias arising from class imbalance in the current dataset.

We experiment with three machine learning models: Support Vector Machines (SVM), Deep Neural Networks (DNN) & Convolutional Neural Networks (CNN).

SVMs are well-understood and widely used, therefore, providing us with a reliable baseline to compare against more complex models. Secondly, SVMs are well-suited for our high dimensional dataset (88 features per file), as they efficiently find the optimal hyperplane that maximizes the margin between classes in such spaces. Moreover, SVM is known to perform well with smaller datasets, considering our positive class data is substantially less compared to the negative class, as it primarily relies on datapoints closer to the hyperplane instead of overall distribution of data. Lastly, SVMs have a range of hyperparameters that can help mitigate overfitting, handle non-linear relationships within data, address class-imbalance by assigning different weights to different classes etc.

We also implement different variants of DNNs and CNNs. We inferred from our trained SVM model that our dataset has non-linear relationships, therefore, applying DNNs and CNNs can be the next reasonable step as they are adept at capturing complex non-linear relationships between our speech features and labels. Additionally, both models automatically capture relevant features from our 88-dimensional data where manually engineering features can pose challenges.

We shall also be exploring the following three data balancing techniques:

1. Class-weighting: this method would assign higher penalties to misclassified samples from the minority class, therefore helping to balance the importance of both ‘No Cold’ and ‘Cold’ classes during training. This method does not generate synthetic data hence maintains integrity of original dataset

2. SMOTE (synthetic minority over-sampling technique): SMOTE will help generate synthetic samples that resemble the minority class. This way, SMOTE helps preserve the underlying distribution of minority class while explicitly addressing class imbalance by inflating samples in this class.

3. ADASYN (adaptive synthetic sampling) tweaks SMOTE by focusing more on generating samples for the minority class that are harder to classify, therefore, addressing class imbalance more effectively. However, both these techniques are computationally expensive and complex in nature.

Our dataset contains significant outliers; therefore, we normalised it to ensure our models are less sensitive to these outliers, resulting in a more stable training process. We implemented SVM, DNN, and CNN models and experimented with data-balancing techniques such as class-weighting, SMOTE, and ADASYN. It is important to note that splitting the training data into training and validation sets after applying these over-sampling techniques can lead to data leakage, thus biasing our models.

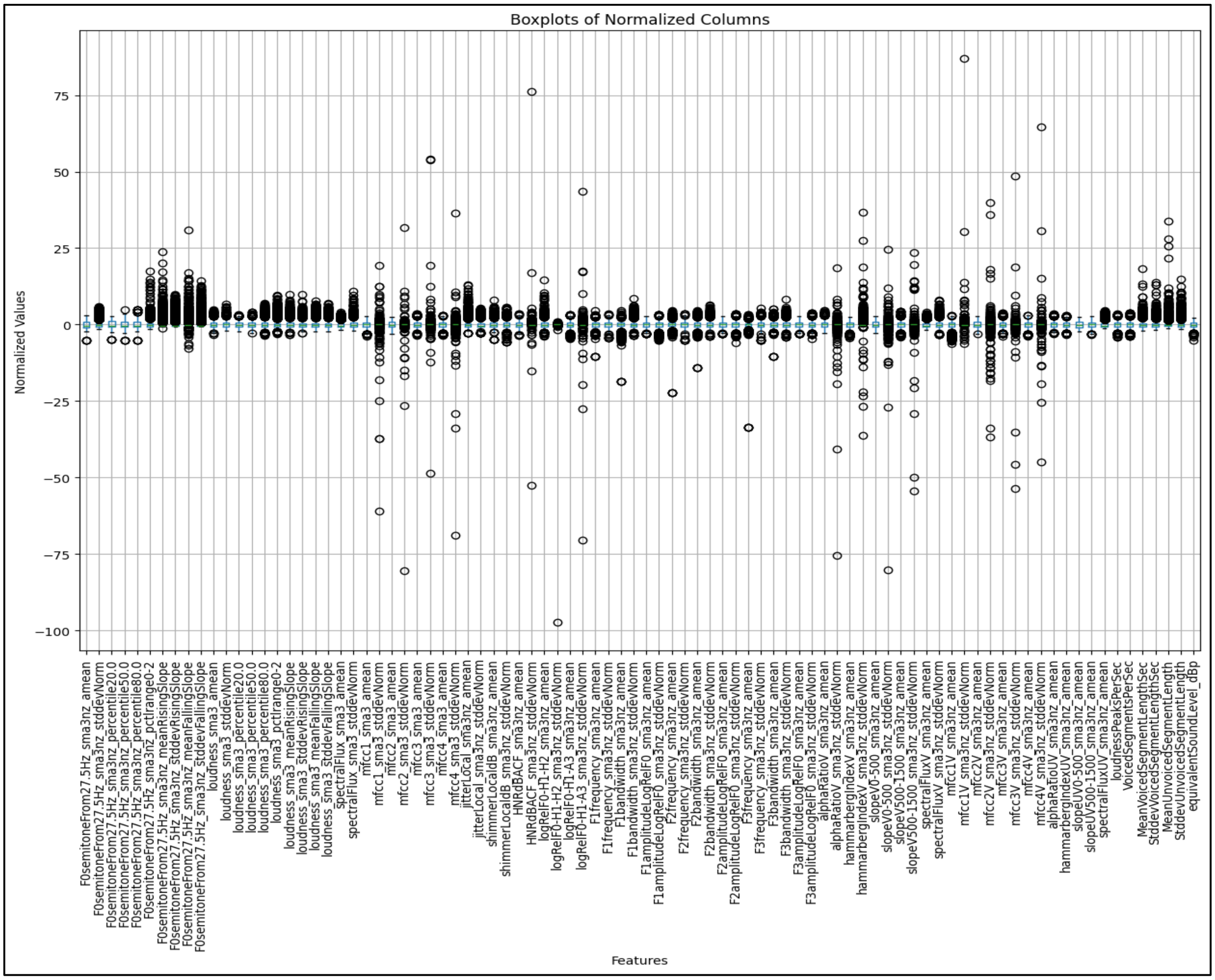

Fig 1. Normalised our data

SVM

For the SVM model, we performed a grid search using cross-validation to select hyperparameters on our unbalanced training set. The RBF SVM achieved the best unweighted average accuracy of 64% on our unbalanced training dataset, indicating that a non-linear kernel captured patterns in the data better than linear kernels. This suggested the presence of non-linear relationships in the data. Once our optimal hyperparameters were selected, we trained our model on the data-balanced training set and then evaluated the performance on the development set, which was unseen during training.

DNN and CNN

For our DNN and CNN models, we split our original training set into training and validation sets before applying over-sampling only to the training set to avoid data leakage. We used the validation set to select optimal hyperparameters. After determining the best hyperparameters, we trained our models on the data-balanced training set and evaluated their performance on the original development dataset.

For our DNN architecture variants, deeper architectures [128,64,32] and [256,128,64] were preferred for their ability to capture more intricate data patterns. ReLU was selected for its ability to handle non-linearities. RMSprop displayed better performance in balanced dataset while Adam was suitable for unbalanced data. Interestingly, all our DNN models preferred 0% dropout rate, indicating that models were not prone to overfitting.

For our CNN architecture variants, the choice of hidden layer configurations [64,32] for unbalanced data and [128,64] for class-weighted and SMOTE data reflected the need to capture varying levels of abstraction in the data. ReLU was consistently selected over Tanh, indicating effectiveness of non-linearity while preventing vanishing gradient issues. Again, no dropout indicates that model was not prone to overfitting. One limitation of our approach is that we did not monitor validation loss during training. If validation loss increases with a 0% dropout rate, it could indicate that the model is not sufficiently complex, potentially resulting in underfitting. This could also explain why both our DNN and CNN models consistently opted for deeper architectures with more hidden layers and neurons. Moreover, due to time constraints, we did not experiment with filters, kernel and pool size and kept them fixed at 32, 3 and 2 respectively.

I assume here that the cost of missing a cold case in the population is very high, therefore, I adjust the threshold for binary classification such that it maximises UAR instead of relying on default threshold of 0.5. Since our data is heavily imbalanced, increasing true positive rate of 'cold' cases came at the cost of increased false positive rate of 'non-cold' cases, i.e., some non-cold cases were misclassified as cold. On ground, this might lead to unnecessary precautions but at least ensures that most cold cases are caught. Therefore, this trade-off prioritised public health by preventing the spread of illness.

Our best performing SVM model was when class-weighting was applied to balance training data with threshold set to 0.63, achieving UAR of 64% on development set. Increasing threshold from default 0.5 significantly increased recall from 30% to 58%, making the model better at identifying actual positives (cold cases) though it also increases false positives.

Despite over-sampling, ADYSYN and SMOTE achieved similar UAR at the cost of increased training time and use of computational resources. Therefore, we select SVM with class weighting as our first final model.

Similarly, our best performing DNN model was with class-weighting and threshold set to 0.99, achieving UAR of 62%. Although very high, it appeared like a reasonable threshold to explore given our DNNs would predict highly skewed probabilities, i.e., either very close to 0 or close to 1 and very few in between (Figure 2). Similar results were observed in CNN model with 0.99 threshold, with UAR maximising at 61% for both class weighting and SMOTE.

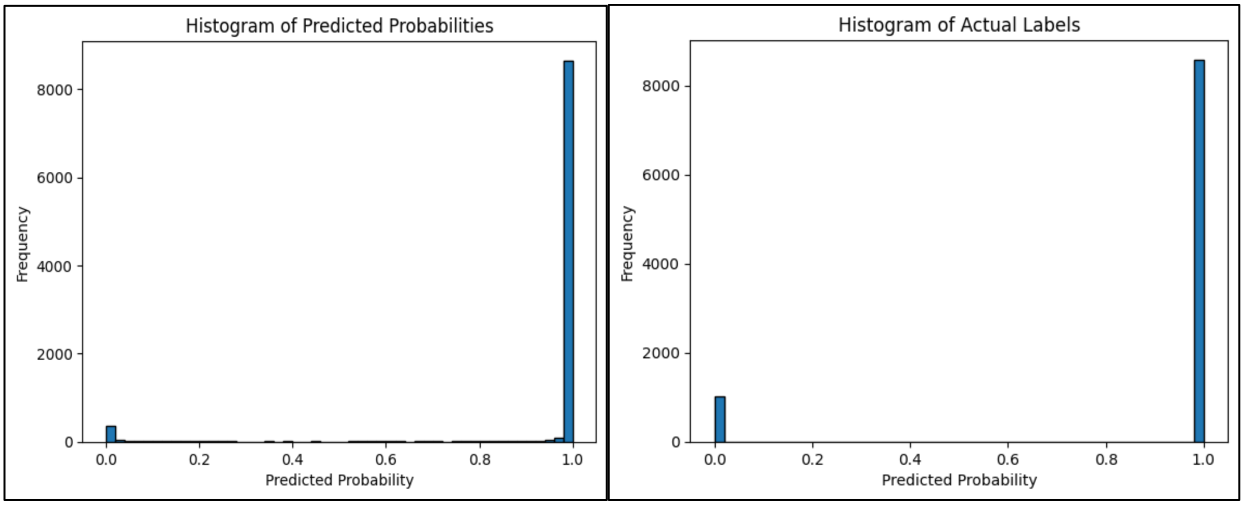

Figure 2. Histogram of predicted probabilities vs histogram of actual probabilities

When evaluating our finalised models on test set, we achieve UAR of 61%, 58% and 53% on SVM, DNN and CNN models respectively. Our DNN and CNN models showed close to 100% on recall but very low precision for cold class, indicating that they are predicting many instances as cold when many of these predictions are, in fact, incorrect. This could be attributed to the high threshold of 0.99 that led to overfitting on the development set. One possible way to improve the performance of our neural network models could be to train them on original imbalanced training data, feed the predicted probabilities to a logistic regression with true labels as target, and then enabling this regression model to learn a function that would map these predicted probabilities to better-calibrated probabilities.

XAI

We have implemented SHAP (SHapley Additive exPlanations) to explain our two models - SVM & DNN. SHAP is model agnostic, i.e., it can be applied to any ML model regardless of its complexity or tyle (linear, non-linear, neural networks etc.). This flexibility has allowed us to use SHAP to interpret and compare both SVM & DNN models effectively. Moreover, SHAP provides global understanding of feature importance across our entire dataset, providing us with the directionality of feature effects. We compared the top predictors identified by SHAP for both models, interpret their directionality and analyse the range of SHAP values for each model, helping us determine how sensitive the model predictions are to the features.

The SHAP plots of the models using 88 speech features showed both similarities and differences for the top 10 list of predictors. Both models shared mfcc3_sma3_amean, logRelF0-H1-A3_sma3nz_amean and loudness_sma3_meanFallingSlope as speech features holding significant predictive power. As 'mfcc3_sma3_amean increases', the probability for participant to have a cold increases. As 'logRelF0-H1-A3_sma3nz_amean' and 'loudness_sma3_meanFallingSlope' increases, the probability for a participant to have a cold decreases. Our DNN model has a smaller SHAP value range (-0.2,0.2) compared to our SVM model (-0.5,0.5) indicating that DNN predictions are less sensitive to changes in our speech features. This was further verified by observing on average a higher impact of speech features on model outcomes using bar plots.

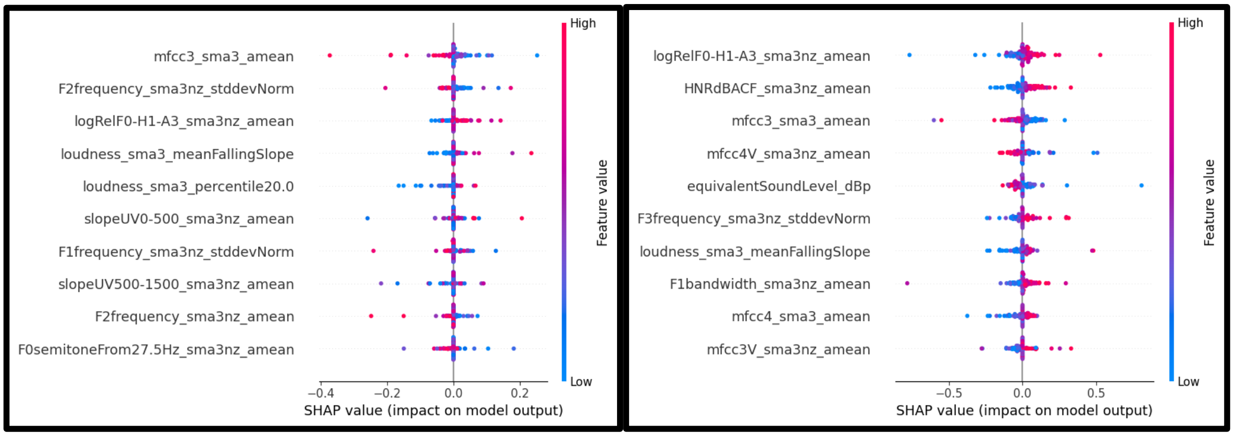

Figure 3. SHAP values for DNN model vs SVM model

GENDER BIAS

Based on previous literature showing gender-specific speech characteristics—such as higher fundamental frequency (F0) and formant frequencies (F1, F2, F3) variability in women, and increased loudness in men (Hardy et al, 2020)—we selected these features to classify data points by gender. K-means clustering approach would be particularly useful when exploring our dataset where distinct patterns related to gender-specific speech traits might exist, as indicated by previous research (Gelfer et al,2005). It operates as an unsupervised learning method, making it suitable for our dataset which lacks explicit gender labels. Therefore, using K-means clustering with 2 clusters, we interpreted cluster centers and observed significant differences in these features, facilitating gender labeling. The clustering process seemed to separate samples based on selected features, indicating distinct speech patterns between genders.

We then partitioned the development set by gender and assessed model performance separately for each group. Across all three models, we noted higher UAR for male-labeled data compared to female-labeled data, suggesting model bias. This observation suggests that certain speech features used for cold/no-cold prediction may be biased towards male speech characteristics. Moreover, the dataset exhibited imbalance with male samples predominating. When specific speech features are more discriminative for males than females, models may struggle to generalize effectively to the underrepresented female class. To address this, we can consider features equally relevant for both genders. Another strategy involves stratifying samples by gender, applying normalization to speech features, and then training the model. This approach helps the model focus on the underlying relationship between speech features and the presence of a cold, rather than picking up gender-specific patterns in raw speech values.

In conclusion, our dataset was highly skewed in favour of 'no cold' class. We attempted to address this by implementing class-weighting, SMOTE and ADYSYN while evaluating model performance using UAR. In terms of balancing, we found little evidence that over-sampling techniques like SMOTE or ADASYN improved the UAR scores. Moving thresholds to optimise UAR improved SVM's performance, however, resulted in overfitting on development set for our neural networks as evidenced by drop in performance on test set.

Symptoms of cold vary considerably across individuals, time and viruses. Therefore, it is possible that the cohort with cold in our dataset is heterogenous. Moreover, if we had more than one speech recording per participant, we could have further controlled for within participant variation, helping us isolate the effect of cold on speech. The model could have benefited with demographic information of participants apart from speech features. This becomes evident when we observe a gender bias affecting the relationship between speech features and our predictions.

REFERENCES

Consider using UAR instead of Accuracy for Imbalanced Classification tasks - Sewade Ogun's Website. (n.d.). Ogunlao.github.io. https://ogunlao.github.io/blog/2021/04/24/consider_uar_accuracy.html

Park, K. (2021, January 15). Voice health during cold season. Premiere Speech & Hearing. Retrieved July 15, 2024, from https://premierespeechhearing.com/voice-health-during-cold-season/#:~:text=When%20you%20have%20a%20cold,muscles%20surrounding%20them%20can%20tighten.

Hardy, T. L. D., Rieger, J. M., Wells, K., & Boliek, C. A. (2020). Acoustic Predictors of Gender Attribution, Masculinity-Femininity, and Vocal Naturalness Ratings Amongst Transgender and Cisgender Speakers. *Journal of voice : official journal of the Voice Foundation*, *34*(2), 300.e11–300.e26. https://doi.org/10.1016/j.jvoice.2018.10.002

Gelfer, M. P., & Mikos, V. A. (2005). The relative contributions of speaking fundamental frequency and formant frequencies to gender identification based on isolated vowels. *Journal of voice : official journal of the Voice Foundation*, *19*(4), 544–554. https://doi.org/10.1016/j.jvoice.2004.10.006