Human Breathprints as Biometric Identifiers: Assessing Uniqueness and Reliability

Michael McCallion
MSc Artificial Intelligence
Ulster University
Belfast, Northern Ireland
mccallion-m15@ulster.ac.uk

Abstract— **Breathprint analysis is emerging as a promising non-invasive biometric identifier, with definite potential to compete with more traditional methods like fingerprint scanning and facial recognition. In this study, two hybrid machine learning models were developed to evaluate the effectiveness of a human 'breathprints' ability to uniquely identify patients, through utilising a dataset of Volatile Organic Compounds (VOC's) collected from patients' breath samples. The first model employed dimensionality reduction through implementing Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA), before applying regression-based feature selection algorithms called Lasso, Ridge and ElasticNet. The second model altered this approach slightly by applying the regression techniques used for feature selection prior to LDA. Both models then had their performance assessed using classifier techniques including Logistic Regression, Support Vector Machine (SVM) and Random Forest. Analysis demonstrated that altering the sequence of feature selection and dimensionality reduction can significantly impact classification accuracy. These results suggest that while breathprints have considerable potential as a biometric identifier, refining the analytical approach is key for optimizing classification outcomes. Further work is needed to fully harness the capabilities of breathprint analysis for reliable and accurate identification of individuals.**

Keywords— Breathprint Analysis, Machine Learning, Volatile Organic Compounds (VOCs), Linear Discriminant Analysis (LDA), Random Forest

I. Introduction

Biometric identification techniques have been at the forefront of security, medical and personal identification systems for many years now. Commonly used techniques include fingerprints, retinal scanning and facial recognition. However, in recent years it has become apparent that the need for more non-invasive techniques has increased. An ideal candidate for this is the use of mass spectrometry of the human breath to identify a unique individual, amongst a group of 'breathprints'. A human 'breathprint' contains compounds known as Volatile Organic Compounds (VOC's). In 1971, Linus Pauling proved that the exhaled human breath contains over 250 substances [1]. With the equipment available to us today, it is possible to identify over 1000 unique substances in the exhaled breath. These VOC's can be used to uniquely identify which subject (patient) the breath belongs to. In this paper, we utilise data processing, cleaning and wrangling techniques, exploratory data analysis, statistical analysis, data visualisation as well as various machine learning algorithms (regression and classification models), to develop a hybrid model capable of successfully identifying an individual simply by running the models developed on their breath or spectral data. Methods used today, including Gas Chromatography-Mass Spectrometry (GC-MS) and Secondary Electrospray Ionisation Mass Spectrometry (SESI-MS) [2], can measure VOC's down to a parts per trillion (ppt) range. When it comes to the actual collection of breath spectral data from an individual, collecting a 'breathprint' is considered much less invasive than the likes of collecting a fingerprint or a retinal scanner. The individual or patient must simply exhale into a sensor. Whereas for both fingerprints and retinal scans, the subject must remain within close proximity to a device, which can cause discomfort for certain patients.

The 'breathprint' can also be used for monitoring metabolic conditions as well as being used for the early detection of diseases. In this paper [3], it is proven that volatile organic compounds contain specific biomarkers that can be indicative towards specific diseases, like diabetes and hyperglycemia. We also see in [4], that VOC's "can hold the key for detecting early metabolic pathway changes in carcinogenesis". These papers evidence that not only is this method of non-invasive biometric collection useful for patient identification, but it also possesses numerous other advantages. Paper [5] documents that VOCs found in the breath, can be indicative of certain metabolic and pathological changes in the body.

Although the research surrounding the human breathprint in relation to uniquely identifying subjects is promising, there is a disconnect in research regarding how reliable the bio-identifier is. This project aims to try to bridge the gap, through the production of a machine learning classifier model, that can identify a patient or individual through a given test sample, with a relative degree of confidence.

II. Literature review

As previously mentioned, biometric idenfitication has been of importance for many years now, with the most commonly used being fingerprint, iris and facial recognition. The use of exhaled human breath for unique identification is an emerging technique that is just now beginning the gain traction. The means of collecting the human breathprint is a non-invasive approach whereby the user exhales into a device which then collects the data or Volatile Organic Compounds (VOC's). These VOC's vary by individual, depedendent on their age, sex, diet and health condition. In [6], it is highlighted how VOC's can provide important insight into the current condition of an inidviduals body, and whether or not they are at risk of any diseases. Through the same logic, a breath print analysis should be able to be used to identify a patient uniquely, from a database of existing patients. In a paper [7], that studied the volatile compounds that exist in the human breath, they grouped VOC's into "chemical functional groups (e.g., methylated hydrocarbons or aldehydes; based on known metabolic and enzymatic pathways)." In the same paper, they were able to achieve a 93% classification accuracy for cancer within individuals, using a Linear Discrimnant Analysis (LDA) and Random Forest classifier model. [8] Shows us that VOC's can be used for their association with numerous diseases. Airway diseases such as asthma and chronic obstructive pulmonary disease (COPD) can be identified within patients whos 'breathprint' contains specific "carbonyl-containing classes (i.e., aldehydes, esters and ketones) and hydrocarbons (i.e., alkanes and alkenes)." Although in the above mentioned paper [8], that VOC's can identify airway dieases, it was mentioned that although these biomarkers do show promise, future work within the area should prioritise the clear reporting of accuracy and multiariate methods, whilst still working on the classification of these volatile metabolites using machine learning algorithms.

In [8], we can see that there has been development of a Convolutional Neural Network (CNN), as well as other neural network models (Multilayer Perceptron and Recurrent Neural Network), that produced an overall accuracy of 97.8% in distinguishing healthy patients from patients that possess lung cancer. [9] This paper found that carcinogenic VOCs exist and there is a relationship between benzene and formaldehyde being present in exhaled breath linking to forms of cancer, including lung and prostate.

Recently, the use machine learning (ML) algorithms within the field of biometrics have been increasing in popularity. These ML techniques focus primarily on pattern recognition, classification and prediction. One of the reasons ML has become integral to the field of biometrics is due to the algorithms' ability to handle datasets that are high-dimensional and complex. In [10], the study delves into and discusses the various ML strategies for improving recognition rates and accuracy in complex datasets. This pattern recognition can be applied to human breath print analysis, as the machine learning algorithms can detect miniscule variations in VOCs, which can be used for differentiation between patients, in turn allowing the patient to be identified amongst others. These AI models can also easily classify biometric data into classes or categories. Using an example related to this paper, classifying exhaled breath data to a specific patient or individual. This is one of the most important factors when it comes to developing the non-invasive technique of breath analysis further. As the machine learning model learns from the data which it has been trained on, it develops establishes an ability to make accurate and specific predictions. Through the analysis of specific trends and relationships within the 'breathprint' data, the model can detect potential health issues during the early stages. Here [11], we can see a study that explored the use of machine learning algorithms to develop a biometric system based on exhaled breath from a human. The system developed utilised ML models to authenticate users, achieving high accuracy rates in confirming these users. Further reviews on the literature surrounding this area can be found in the supplementary materials.

III. PROBLEM SCENARIO

In the area of biometrics, fingerprint scanning has often been perceived as the best available option to subjects or individuals due to how reliable and secure of a method it is. However, as the requirements for healthcare technology evolves within the field of digital health, there is a high demand for the development and implementation of a non-invasive biometric identification method that offers the same (if not better) levels of accuracy and reliability.

One such example of a non-invasive method for this is the use of human 'breathprints' as a biometric identifier. Volatile Organic Compounds (VOC's) that can be analysed within the human breath can be the key to proving if exhaled human breath could be as uniquely identifiable as the likes of a fingerprint or iris. It holds the potential to revolutionise biometric identification, particularly in the field of digital health. This project aims to discover whether human 'breathprints' could provide a reliable, unique and accurate biometric identifier, comparable to that of fingerprint or facial recognition.

In healthcare, the demand for non-invasive monitoring tools is increasing in demand, with breathprint analysis offering a means to seamlessly identify patients whilst providing valuable observations on their health status. Granted that 'breathprints' could successfully be discovered to be as powerful as fingerprints, they could be used in applications such as a secure and safe way to identify patients and gain access to their medical records. More importantly, before it is adopted as a tool for identification, we must address the fundamental question of their uniqueness and reliability.

Three of the main challenges in this project are deemed to be:

1. *Data Variability* – Datasets containing varying VOC values in the form of continuous variables poses the issue of the inherent variability within the dataset. The VOC's present within a human breath constantly fluctuate and change throughout their lifetime. Dependent on several factors including age, diet, health conditions and exposure to certain environments. Naturally, the variability of such data elicits questions relating to the stability and uniqueness of breathprints over time. The overall question remains – can a breathprint consistently identify a subject or will the variations within the data undermine the reliability of the tool.
2. *Small Sample Sizes* – An additional challenge is how widely available large, labelled datasets are, that capture the entire range of breathprint variability. As it is seen as a new field of study, with most research occurring in the last 4 years, there is a lack of extensive datasets containing exhaled breath information. The use of small datasets and in turn, small training and test sets applied to machine learning models, can result in overfitting. This means that the models perform well on training data, but fail to perform as well on new, unseen data when the models have been deployed.
3. *Overfitting & Model Complexity* – As breathprint data is extremely complex, machine learning models are used to analyse the data and then classify it accurately. Again, due to the complex nature of the data, the models can be at risk of overfitting. To mitigate the risk, cross-validation, regularization and data augmentation techniques are utilised.

Breathprint analysis is said to still be in its exploratory phase. The majority of studies are limited by small sample sizes with high data variability.

The primary objective of this project is to determine whether human 'breathprint' can be as uniquely identifiable as human fingerprints. More specifically, the study aims to:

1. *Assess uniqueness* – Investigate uniqueness of 'breathprints' by analysing the variability of VOC's patterns including their inter-individual and intra-individual relationships.
2. *Mitigate data variability* – Apply robust and effective pre-processing techniques to normalise breathprint data.
3. *Address small sample sizes* – Use cross-validation and data augmentation to improve the generalisability of the models, despite the challenges posed by limited datasets.
4. *Prevent overfitting* – Implement regularisation methods and validate model performance.
5. *Evaluate classifier models* – Implement classification models and provide an evaluation for each model through understanding their accuracy, precision, recall and f1-score metrics.

Determining if breathprints can be utilised as a unique biometric identifier, comparable to fingerprints, is crucial for advancing the field of non-invasive biometric identification. If successful, the research could be the steppingstones for breathprint to be used in several different applications (e.g., secure patient identification in healthcare). By addressing the challenges of data variability, small sample sizes and model overfitting, this project aims to provide an analysis of the human breathprints' potential as a biometric tool.

IV. METHODOLOGY

The following section outlines the methodological approach that was planned to be used to analyse breathprint data for biometric identification. The process will involve multiple techniques for statistical analysis as well as for machine learning. These include Kruskal-Wallis testing, Bonferroni correction, Principal Component Analysis (PCA). The machine learning models implemented regression algorithms called Ridge, Lasso and ElasticNet for feature selection. Models known as Logistic Regression, Support Vector Machines (SVM) and Random Forests were used for classifying the test data after the models have been trained. The overall goal, as previously mentioned, is to identify the most important features from the dataset and then classify the test VOC samples to a specific individual or patient.

A. Data Collection

The dataset used was generously provided by Martinez-Lozano Sinues, P. et al. [2], the authors of study performed in Switzerland. They have granted permission for this to be used in the study. The original study involved eleven subject (6 males, 5 females) with a heterogeneous ethnic background. Over nine days, breath samples were collected at specific times. Food and drink intake were controlled prior to measurements being taken. Breath samples were analysed using a quadrupole time-of-flight mass spectrometer (Q-TOF), employing a secondary electrospray ionisation (SESI) to identify the VOCs. Further details of the data collection including ethical considerations are documented in the original study [2].

B. Data Splitting

The dataset was divided into two subsets:

- Training set – This comprised of 80% of the original dataset.
- Test set – This comprised of 20% of the original dataset.

This train-test split allowed for the machine learning models to be trained on one subset of data (training set) and then to best tested on a completely separate, unseen dataset (test set), to evaluate the model's performance and thus providing evaluation metrics.

C. Scaling the Data

As a result of the varying scales of VOC's, it was crucial to standardise the dataset before proceeding with dimensionality reduction. To achieve this, the 'StandardScaler()' function was used to transform the data. This meant each feature within the dataset had a mean value of zero, with a standard deviation of one. This means that all features contribute equally to the analysis, especially when working with sensitive techniques like PCA and LDA. This process was key in preparing the data for later steps in the analysis. Scaling helped to reduce the impact of outliers and ensured the data remained on a consistent scale. It helped make sure that the variability found within the dataset was due to unseen patterns within the data and not differences in the scale of features.

D. Principal Component Analysis (PCA)

Moreover, the next stage in the process was to apply Principal Component Analysis (PCA) to the statistically significant features that had been extracted through the previous stages mentioned. The purpose of this was to further reduce the size of the dataset whilst retaining a specific degree of variance. In this study, 99.99% of variance was retained whilst performing PCA. PCA is known as a linear dimensionality reduction technique. The goal of applying PCA is to transform the data into a set of principal components [20]. These new principal components (PC's) are ranked by the amount of variance which they explain. The cumulative value of the variance of all principal components will come to a total of 99.99%.

As PCA was applied whilst retaining 99.99%, it assured that the most important features remained whilst eliminating any redundant or noisy data that remained [19]. As the dataset has now been greatly reduced, it reduces the computational resources required for machine learning models and helps eliminate the threat of these models' overfitting. Before the PCA was applied, a standard scaler was used in order to standardise the data. This is due to PCA being sensitive towards the scale of variables within the dataset [20]. A variance retention of 99.99% allowed for the maintenance of the majority of information from the original dataset, whilst effectively reducing the dimensionality of the dataset.

E. Linear Discriminant Analysis (LDA)

In the 1st iteration of the model, Linear Discriminant Analysis was applied after PCA, prior to the feature selection regression algorithms being implemented. In the 2nd iteration, LDA was applied after the feature selection process. The goal of LDA is to maximise separability between classes (patients). LDA is a supervised technique that finds linear combinations of features which best separate classes [21]. Through doing so, LDA was able to isolate the most discriminative features within the dataset.

The application of LDA produced a refined dataset which contained the most relevant components for classification. Through reducing the features to only those that were most statistically significant discriminants, LDA enhanced the classifier model's ability to successfully identify patients. This step allowed for the final analysis to focus on features which possessed the most significant biomarkers for patient identification.

F. *Feature Selection using Regression models*

Lasso Regression

One of the regression algorithms implemented for feature selection is known as Lasso (Least Absolute Shrinkage Operator). This linear regression technique possesses an L1 regularisation penalty. This penalty shrinks a number of the models less important coefficients to zero [22]. This technique was particularly useful within this project as it helped to identify the most relevant VOC's by eliminating features with little or no contribution to the model. As Lasso selects a specific subset of features, it improves the interpretability of the model.

$$L_{lasso}(\hat{\beta}) = \sum_{i=1}^{n}(y_i - x_i^T\hat{\beta})^2 + \lambda\sum_{j=1}^{m}|\hat{\beta}_j|$$

Figure 1 - Lasso Regression Formula [29]

Ridge Regression

A second linear regression method known as Ridge regression was applied. This model uses an L2 regularisation penalty, which discourages large coefficients, but does not shrink them to zero (like Lasso does) [23]. Ridge regression is used to determine whether regularisation without sparsity would return a more valuable feature selection. Ridge keeps all features. However, it penalises its coefficients so to reduce multicollinearity and improve model generalisation [24].

$$SSE_{L_2} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda\sum_{j=1}^{P}\beta_j^2$$

Figure 2 - Ridge Regression Formula [30]

ElasticNet Regression

The ElasticNet regression model combines L1 and L2 regularisation to cater towards a more balanced approach to feature selection [25]. The use of ElasticNet was crucial in this project as it allowed for a compromise between the techniques used in both Lasso and Ridge. ElasticNet is found to be effective when the dataset contains highly correlated features and ensuring relevant features were selected whilst penalizing less significant features.

$$\text{Loss} = \Sigma^n_{i=1}(y_i - \hat{y}_i)^2 + \alpha_1\Sigma^n_{i=1}|w_i| + \alpha_2\Sigma^n_{i=1}w_i^2$$

Figure 3 - ElasticNet Regression Formula [31]

G. *Classification Models*

Logistic Regression

Logistic regression was applied to classify the VOC test samples. This model predicts the probability of a binary outcome [26]. This model's purpose was to provide a baseline comparison against other more complex models employed. Although this model is seen as a basic or simple technique, it was found to be a useful tool in this study.

Random Forest

This model was also implemented to classify the VOC test samples. It is known as an ensemble learning method [27] that constructs several decision trees, then merging them to produce a single, accurate and stable prediction. Random forest is extremely effective at handling highly dimensional data. It averages the prediction values of multiple decision trees and in turn reduces variance whilst improving generalisation.

Support Vector Machines (SVM)

This model works by finding the optimal hyperplane that separates data into different classes [27], maximizing the margin between closest points of different classes (i.e., support vectors). The SVM's models' ability to deal with non-linear relationships made it a strong choice for 'breathprint' classifiers.

H. *Model Evaluation & Validation*

10-fold cross validation was performed on the training data to ensure robustness and generalisability of models. This process involves splitting the data into 10 subsets, with the model being trained on 9 of them and then validated on the final remaining subset. This process is iterated 10 times in total, with each subset serving as the validation set once. This cross-validation technique was pivotal in preventing overfitting and ensuring consistent model performance. The results from each fold were averaged, providing a more accurate and reliable estimate of the model's accuracy.

Performance Metrics

Model performance was evaluated using accuracy, precision, recall and f1-score metrics. The metrics provided a comprehensive assessment of the models' ability to correctly classify VOC samples and assign them to the correct subject. Results addressed in next section.

V. RESULTS

The first approach to analyse the dataset was implementing a hybrid model which utilised regression and classification machine learning algorithms, after dimensionality reduction techniques had been performed. Firstly, scaling of the data took place using a StandardScaler so that each feature within the dataset contributed equally to the data analysis. PCA was applied to the scaled data, with a variance retention of 95%. The number of features was reduced significantly with the PCA train set shape being "154, 13" and the PCA test set shape being "39, 13". In total, the 193 patients' samples were reduced to a total of 13 principal components. This is a significant reduction that still encapsulates the vast majority (95%) of original variance.

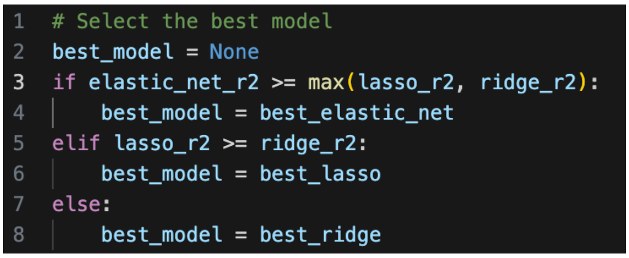
Next, Linear Discriminant Analysis (LDA) was applied to the previously transformed PCA dataset. This supervised dimensionality reduction technique aims to maximise the separation between classes. The implementation of LDA further reduced the dataset to 193, 10. Meaning that the LDA algorithm identified 10 discriminant components to be the most significant in distinguishing between classes in the dataset. Note, that all discriminant components were retained to preserve full discriminatory power of the model.

Three regression models were employed: Lasso, Ridge & ElasticNet. Each model applied a penalty to less relevant features and in turn enhanced the model's interpretability and reduced the chances of overfitting.
The performance of each regression model was evaluated on their R^2 score. R^2 score indicates the proportion of variance in the dependent variable that is predictable from the independent variables [28]. The R^2 scores for each were as follows:

R^2 scores of Regression Models (1st Iteration)	
Regression Model	***R^2 Score***
Lasso	-0.0195
Ridge	-0.0106
ElasticNet	**0.0016**

Table 1 - R^2 scores for Regression Models

As seen above, the ElasticNet model was the best performer. Although, the R^2 value for each of the models are low, meaning they all struggled to explain the variance in the dataset.



```python
# Select the best model
best_model = None
if elastic_net_r2 >= max(lasso_r2, ridge_r2):
    best_model = best_elastic_net
elif lasso_r2 >= ridge_r2:
    best_model = best_lasso
else:
    best_model = best_ridge
```

Figure 4 - Else-if statement used to automatically select best model

An else-if statement (figure 1) was then implemented which chose the highest performing model based on which had the highest R^2 score. The parameters of the best performing ElasticNet model along with the other two Lasso and Ridge models, including their optimal hyperparameters used can be seen following:

Optimal Models for Feature Selection (1st Iteration)			
Model Names	***Alpha value***	***L1_ratio***	***R^2 score***
Lasso	0.3449	N/A	0.0195
Ridge	**9.8699**	N/A	**0.1056**
ElasticNet	0.8859	0.1960	0.0016

Table 2 - Regression models and their optimised parameters (1st Iteration)

Using the selected features from the ElasticNet model, three different classification models were applied to classify the test VOC samples to specific individuals. The 3 models used were:

1. Logistic Regression
2. Support Vector Machines (SVM's)
3. Random Forest

The performance of each model along with their evaluation metrics can be seen in the table below:

Classification Methods Evaluation Metrics (1st Iteration)				
Model Names	***Accuracy***	***Precision***	***Recall***	***F1-score***
Logistic Regression	25.64	28.61	25.64	24.46
SVM	30.77	**40.72**	30.77	**33.03**
Random Forest	**33.33**	32.46	**33.33**	32.01

Table 3 - Evaluation Metrics for Classifier Models (1st Iteration)

The logistic regression model achieved sub- standard results, demonstrating low scores across all metrics. An accuracy of 25.64% shows that the model incorrectly classified the vast majority of test samples.

The SVM model outperformed that of the Logistic Regression model, achieving both a higher accuracy and precision (30.77% accuracy and 40.72% precision). The increase in the f1-score (33.03%) evidences a greater balance between precision and recall. This implies that SVM experienced a greater success at distinguishing between classes (patients).

The Random Forest model achieved the highest accuracy of 35.90%. This model achieved an f1-score of 33.15%, like that of the SVM model. The performance of this random forest model shows that it was the best suited to handling the complexity of the VOC data, meaning it is the most promising classifier of these tested.

The second model iterated was slightly different from the first model. In this model, Principal Component Analysis (PCA) is not used. The regression algorithms for feature selection are run first, then Linear Discriminant Analysis (LDA) is applied to the selected features. The aim of this was to refine the feature selection process and then transform the dataset into discriminant components by utilising the dimensionality reduction technique. The results from this model can be found following.

As per the previous model, the dataset was scaled before undergoing feature selection. Each model had the same randomized search hyperparameters as in the first iteration. The R^2 score for each can be found below:

R^2 scores of Regression Models (2nd Iteration)	
Regression Model	***R^2 Score***
Lasso	-0.7125
Ridge	**-0.0171**
ElasticNet	-0.8987

Table 4 - R^2 scores for Regression Models (2nd Iteration)

Although all R^2 scores are negative, the Ridge regression model was the best performer. Although this score is not ideal, it was significantly better than that of the Lasso and ElasticNet regression models. These negative values once again suggest that all three models struggled to explain variance in the dataset adequately.

The same else-if statement (figure 1) was used to automatically determine the best model (based on R^2 scores). The models along with its optimal parameters and R^2 score can be seen below.

Optimal Models for Feature Selection (2nd Iteration)			
Model Names	***Alpha value***	***L1_ratio***	***R^2 score***
Lasso	0.0553	N/A	-0.7125

Ridge	0.0553	N/A	**-0.0171**
ElasticNet	**0.0696**	0.5107	-0.8987

Table 5 - Regression Models and their Optimised Parameters (2nd Iteration)

Following feature selection using regression, LDA was then applied to the new dataset to reduce dimensionality. The process of LDA transformed the data into 10 linear discriminants. This reduced the dataset to a shape of 193, 10, the same as that found in the first iteration, whereby the dataset contained 10 discriminant components in total.

This proves that the LDA was successful in reducing the dataset whilst retaining the discriminatory power required for classification. The same 3 models were applied, with the following table displaying their metrics.

Classification Methods Evaluation Metrics (2nd Iteration)

Model Names	*Accuracy*	*Precision*	*Recall*	*F1-score*
Logistic Regression	**56.41**	**65.81**	**56.41**	**55.99**
SVM	41.03	61.05	41.03	44.74
Random Forest	38.46	32.46	33.33	32.01

Table 6 - Evaluation Metrics for Classifier Models (2nd Iteration)

When comparing the Logistic Regression model to the previous iteration, it has shown significant improvement - across all metrics. Particularly with accuracy, where it has increased from 25.64% to 56.41%. This shows how performing feature selection before LDA and removing PCA altogether is beneficial. From the results, it is clear that by applying the regression models in this fashion, it resulted in a significant increase in the model's ability to classify data.

The SVM model also displayed a noticeable improvement across all metrics, with the two main improvements being accuracy (increased from 30.77% to 41.03%) and recall (increased from 40.72% to 61.05%). This proves that the SVM model also benefitted from the alternative approach.

The Random Forest classifier also demonstrated an improvement in performance, but to a lesser extent than that of the latter two models. The accuracy increased from 33.33% to 38.46% and the precision increased from 32.48% to 54.1%.

A table showing feature importance based on the logistic regression models coefficients can be seen below:

Feature Importance based on Logistic Regression coefficients (2nd Iteration)

Linear Discriminant Component	*Value*
LD5	0.6407
LD7	0.2077
LD4	0.1217
LD9	0.0776
LD10	-0.0875
LD2	-0.1034
LD8	-0.4617
LD1	-0.5926
LD6	-0.6388
LD3	-0.7191

Table 7 - Logistic Regression Coefficients (2nd Iteration)

The positive coefficients indicate that an increase in the corresponding linear discriminant contributes positively to the likelihood of assigning a sample to a particular class. Contrastingly, each negative coefficient suggests the opposite.

The most influential discriminant component was LDS, capturing a coefficient of 0.6407. This implies that this specific component captures a great amount of discriminatory information. Conversely, LD3 and LD6 contributed negatively, meaning the models may have struggled with classification due to the presence of these features.

VI. DISCUSSION

Both iterations of models have displayed insights into how effective different approaches to feature selection and dimensionality reduction are when classifying the human breathprint data. Through the comparison of the metrics across these models, it is apparent that certain approaches impact accuracy and reliability of models.

In the first iteration of the model, where PCA and LDA were applied first, before the regression models, the Ridge regression model outperformed Lasso and ElasticNet with an R^2 score of -0.0106 (table 1). The result shows that whilst Ridge was somewhat effective the model's overall ability to understand the variance of the data was limited. Although ElasticNet (which balances L1 and L2 regularisation) demonstrated a positive R^2 score of 0.0016 (table 1) it was not a strong enough of a score to justify using it, given its slightly better score.

Contrastingly, during the 2nd iteration (where regression models were applied before LDA), the R^2 values decreased across each of the 3 models used, in particular with Lasso and ElasticNet. Lasso exhibited a score of -0.7125 and ElasticNet with -0.8987 (Table 4). In this iteration, the Ridge regression model also saw its R^2 score drop to -0.0171. One possible reason for this may be due to reversing the feature selection and dimensionality reduction processes. In doing so, it may have resulted in the retention of more noise within the dataset, in turn reducing the effectiveness of the regression techniques used for feature selection.

The difference in R^2 scores highlight the importance of the sequence in which the aforementioned techniques were applied. The results show applying both PCA and LDA may be a better approach by allowing the regression models to focus on more prominent or important features. This could be because the LDA maximises separability between classes (patients). Thus, resulting in a better refined dataset for classification.

Among all the models test in the first iteration, the Random Forest model achieved the highest accuracy score of 33.33%, with a recall score of 33.33% also. These metrics would be considered as low, evidencing that the model struggled to classify the data. This may be due to the complexity and high dimensionality of the dataset.

The Logistic Regression model used, which is usually seen as easier to understand and less susceptible to overfitting, demonstrated a poor performance in the first iteration with an accuracy of 25.54% and an f1-score of 24.46%. This may be due to Logistic Regression being a linear model, meaning it may well have struggled to capture complexities that exist within the data. However, the 2nd iteration of the model managed to display a significant improvement in the Logistic Regression models performance. In this iteration, the modes performance increased to 56.41%, with its f1-score increasing to 55.99%

(Table 6). This shows that by applying feature selection first, then LDA after, it helped to boost the Logistic Regressions scores by reducing the dataset to its most predictive features.

The Random Forest model also displayed a slight improvement from the first iteration to the second iteration, with the accuracy score increasing from 33.33% to 38.46%. This again implies that the Random Forest model performs better on a dataset which was reduced using feature selection first, then the classes were maximally separated using LDA.

Moreover, the SVM models performance increased from 1st iteration to 2nd iteration, with accuracy and prediction scores increasing to 41.03% (from 30.77%) and 66.05% (from 40.72%) respectively.

Table 7 displays feature importance based on Linear Regression coefficients. It provides insights into which linear discriminants were the most influential in the classification task. 'LD5' had the highest positive impact (a value of 0.6407) on the models' predictions. This made LD5 the most significant feature in contributing towards predictions. It implies that LD5 captures a vital aspect of the data which holds a strong relationship with the class labels.

In the contrary, 'LD3' has the greatest negative impact (-0.7197). This may indicate that the discriminant component holds features which are negatively correlated towards the classification outcome. The data possessing both strongly positive and negative coefficients demonstrates the complexity of the dataset, with certain features contributing positively and others negatively.

It can be assumed that the model relied on a balance of features that had positive and negative impacts. This reinstates the importance of the dataset possessing a diverse set of discriminants. However, as the majority of the discriminant component values were relatively small, it shows that no single feature dominated the classification task. This reinforces the need for an effective combination of features.

Comparing both iterations shows that some of the adjustments to the methodology led to improvements (especially with the Logistic Regression model), but the overall performance of all 3 models in both iterations remained modest. The key point to recognize here is that the order in which the dimensionality reduction and feature selection techniques are performed can greatly impact model performance.

Future work should focus on further refining these techniques by exploring alternative or additional techniques to implement. More sophisticated regularisation techniques could be employed alongside non-linear dimensionality reduction processes like t-Distributed Stochastic Neighbour Embedding (t-SNE) or Uniform Manifold Approximation and Projection (UMAP).

Given the variability in model performance, further analysis could look into ensemble methods that combine the strength of multiple models to achieve more robust predictions.

The analysis of feature importance reinforces the need to delve into specific characteristics of breathprint data, that contributes the most to classification.

VII. Conclusion

This subsection of the study aimed to test the feasibility of breathprint analysis as a non-invasive biometric identifier, with an emphasis on the application of machine learning methodologies for feature selection and classification. Two hybrid models allowed deep insights into the optimal sequencing of the dimensionality reduction and feature selection process. The first one, applying PCA and LDA before feature selection, provided modest classification accuracy, especially with the case of Random Forests. The 2nd model reversed this process, conducting feature selection for LDA, and performance of all three models was enhanced. The present study gave some evidence that methodological sequencing would help in improving the discriminating power of VOC patterns; on the other hand, the enhancements indicate problems in the classification of high-dimensional and complex data. Future work should consider other methods in the dimensionality reduction technique that would be a better alternative to PCA, and also look at more advanced ensemble techniques in an effort to make better the accuracy of breathprint identification using machine learning techniques.

Acknowledgments

I would like to thank Dr. Omar Nibouche, my project supervisor, for his guidance and support throughout the entirety of this project.

References

[1] Pauling, L., Robinson, A. B., Teranishit, R., & Cary, P. (1971). Quantitative Analysis of Urine Vapor and Breath by Gas-Liquid Partition Chromatography (orthomolecular medicine/vitamins/controlled diet) (Vol. 68, Issue 10). https://www.pnas.org

[2] Martinez-Lozano Sinues, P., Kohler, M., & Zenobi, R. (2013). Human Breath Analysis May Support the Existence of Individual Metabolic Phenotypes. PLoS ONE, 8(4). https://doi.org/10.1371/journal.pone.0059909

[3] Kaloumenou, M., Skotadis, E., Lagopati, N., Efstathopoulos, E., & Tsoukalas, D. (2022). Breath Analysis: A Promising Tool for Disease Diagnosis—The Role of Sensors. In Sensors (Vol. 22, Issue 3). MDPI. https://doi.org/10.3390/s22031238

[4] Peled, N., Fuchs, V., Kestenbaum, E. H., Oscar, E., & Bitran, R. (2021). An update on the use of exhaled breath analysis for the early detection of lung cancer. In Lung Cancer: Targets and Therapy (Vol. 12, pp. 81–92). Dove Medical Press Ltd. https://doi.org/10.2147/LCTT.S320493

[5] Peled, N., Fuchs, V., Kestenbaum, E. H., Oscar, E., & Bitran, R. (2021). An update on the use of exhaled breath analysis for the early detection of lung cancer. In Lung Cancer: Targets and Therapy (Vol. 12, pp. 81–92). Dove Medical Press Ltd. https://doi.org/10.2147/LCTT.S320493

[6] Das, S., Pal, S., & Mitra, M. (2016). Significance of Exhaled Breath Test in Clinical Diagnosis: A Special Focus on the Detection of Diabetes Mellitus. In Journal of Medical and Biological Engineering (Vol. 36, Issue 5, pp. 605–624). Springer Berlin Heidelberg. https://doi.org/10.1007/s40846-016-0164-6

[7] Issitt, T., Wiggins, L., Veysey, M., Sweeney, S. T., Brackenbury, W. J., & Redeker, K. (2022). Volatile

compounds in human breath: Critical review and meta-analysis. In Journal of Breath Research (Vol. 16, Issue 2). IOP Publishing Ltd. https://doi.org/10.1088/1752-7163/ac5230

[8] Ibrahim, W., Natarajan, S., Wilde, M., Cordell, R., Monks, P. S., Greening, N., Brightling, C. E., Evans, R., & Siddiqui, S. (2021). A systematic review of the diagnostic accuracy of volatile organic compounds in airway diseases and their relation to markers of type-2 inflammation. ERJ Open Research, 7(3). https://doi.org/10.1183/23120541.00030-2021

[9] Lee, B., Lee, J., Lee, J. O., Hwang, Y., Bahn, H. K., Park, I., Jheon, S., & Lee, D. S. (2024). Breath analysis system with convolutional neural network (CNN) for early detection of lung cancer. Sensors and Actuators B: Chemical, 409. https://doi.org/10.1016/j.snb.2024.135578

[9] Moura, P. C., Raposo, M., & Vassilenko, V. (2023). Breath volatile organic compounds (VOCs) as biomarkers for the diagnosis of pathological conditions: A review. In Biomedical Journal (Vol. 46, Issue 4). Elsevier B.V. https://doi.org/10.1016/j.bj.2023.100623

[10] Singh, C. (2023). Machine Learning in Pattern Recognition. European Journal of Engineering and Technology Research, 8(2), 63–68. https://doi.org/10.24018/ejeng.2023.8.2.3025

[11] Karunanethy, M., Tripathi, R., Panchagnula, M. v., & Rengaswamy, R. (2024). User authentication system based on human exhaled breath physics. PLoS ONE, 19(4 April). https://doi.org/10.1371/journal.pone.0301971

[12] Li, W., Xu, J., Yang, W., Liu, F., Zhou, H., & Yan, Z. (2024). Approach and application of extracting matching features from E-nose signals for AI tasks. Biomedical Signal Processing and Control, 90. https://doi.org/10.1016/j.bspc.2023.105869

[13] Pegoraro, J. A., Lavault, S., Wattiez, N., Similowski, T., Gonzalez-Bermejo, J., & Birmelé, E. (2021). Machine-learning based feature selection for a non-invasive breathing change detection. BioData Mining, 14(1). https://doi.org/10.1186/s13040-021-00265-8

[14] Suresh, K. C., Prabha, R., Hemavathy, N., Sivarajeswari, S., Gokulakrishnan, D., & Jagadeesh kumar, M. (2022). A Machine Learning Approach for Human Breath Diagnosis with Soft Sensors. Computers and Electrical Engineering, 100. https://doi.org/10.1016/j.compeleceng.2022.107945

[15] de Vincentis, A., Pennazza, G., Santonico, M., Vespasiani-Gentilucci, U., Galati, G., Gallo, P., Vernile, C., Pedone, C., Antonelli Incalzi, R., & Picardi, A. (2016). Breath-print analysis by e-nose for classifying and monitoring chronic liver disease: A proof-of-concept study. Scientific Reports, 6. https://doi.org/10.1038/srep25337

[16] IBM, "What is Overfitting? | IBM," *www.ibm.com*, 2024. https://www.ibm.com/topics/overfitting

[17] Fi, M. O., & Garriga, G. C. (2010). Permutation Tests for Studying Classifier Performance Markus Ojala. In Journal of Machine Learning Research (Vol. 11).

[18] "Bias and Variance in Machine Learning: An In Depth Explanation," *Simplilearn.com*. https://www.simplilearn.com/tutorials/machine-learning-tutorial/bias-and-variance

[19] N. Pudjihartono, T. Fadason, A. W. Kempa-Liehr, and J. M. O'Sullivan, "A Review of Feature Selection Methods for Machine Learning-Based Disease Risk Prediction," *Frontiers in Bioinformatics*, vol. 2, Jun. 2022, doi: https://doi.org/10.3389/fbinf.2022.927312.

[20] A. Biswal, "Principal Component Analysis in Machine Learning | Simplilearn," *Simplilearn.com*, Apr. 24, 2023. https://www.simplilearn.com/tutorials/machine-learning-tutorial/principal-component-analysis

[21] Raman, "ML | Linear Discriminant Analysis," *GeeksforGeeks*, May 03, 2019. https://www.geeksforgeeks.org/ml-linear-discriminant-analysis/

[22] Hesamian, G., Johannssen, A., & Chukhrova, N. (2024). An explainable fused lasso regression model for handling high-dimensional fuzzy data. *Journal of Computational and Applied Mathematics*, *441*. https://doi.org/10.1016/j.cam.2023.115721

[23] Nakatsu, Robbie T.. "Validation of machine learning ridge regression models using Monte Carlo, bootstrap, and variations in cross-validation" *Journal of Intelligent Systems*, vol. 32, no. 1, 2023, pp. 20220224. https://doi.org/10.1515/jisys-2022-0224

[24] C. M. Bishop, *Pattern Recognition and Machine Learning*. 2023.

[25] J. Kenneth Tay, B. Narasimhan, and T. Hastie, "Elastic Net Regularization Paths for All Generalized Linear Models," *Journal of Statistical Software*, vol. 106, no. 1, Jan. 2023, doi: https://doi.org/10.18637/jss.v106.i01.

[26] S. J. Russell and P. Norvig, *Artificial intelligence : a modern approach*. London: Pearson, 2021.

[27] P. A. Flach, *Machine learning : the art and science of algorithms that make sense of data*. Cambridge: Cambridge University Press, 2012.

[28] S. Raschka and V. Mirjalili, *Python machine learning : machine learning and deep learning with python, scikit-learn, and tensorflow 2*. Birmingham: Packt Publishing, Limited, 2019.

[29] Y. Gupta, "Lasso Regression — In 'Simple' English," *Data Science Simplified*, Sep. 29, 2022. https://medium.com/dssimplified/lasso-regression-in-simple-english-8558e354781c

[30] K. Kargın, "Ridge Regression Fundamentals and Modeling in Python," *Medium*, Apr. 17, 2021. https://keremkargin.medium.com/ridge-regression-fundamentals-and-modeling-in-python-bb56f4301f62

[31] Arun Addagatla, "Regularization Techniques in Machine Learning - Arun Addagatla - Medium," *Medium*, Apr. 20, 2021. https://arunaddagatla.medium.com/regularization-techniques-bbc0abfc5fe (accessed Aug. 25, 2024).