A Novel Mixed Effects Random Forest Approach for Predicting Dairy Cattle Methane Emissions

A Novel Mixed Effects Random Forest Approach for Predicting Dairy Cattle Methane Emissions

Stephen Ross
School of Computing, Ulster University,
Belfast, Northern Ireland
ross-s24@ulster.ac.uk

Haiying Wang
School of Computing, Ulster University,
Belfast, Northern Ireland
hy.wang@ulster.ac.uk

Huiru Zheng*
School of Computing, Ulster University,
Belfast, Northern Ireland
h.zheng@ulster.ac.uk

Tianhai Yan
Ruminant Nutrition Unit, Agri Food and Biosciences Institiute,
Hillsborough, Northern Ireland
tianhai.yan@afbini.gov.uk

Masoud Shirali
Ruminant Genetics Unit, Agri Food and Biosciences Institiute,
Belfast, Northern Ireland
masoud.shirali@afbni.gov.uk

Abstract— Methane (CH_4) emissions produced by dairy cattle (DC) are a key contributor to global warming. To assess the effectiveness of strategies designed to mitigate CH_4 emissions, complex and expensive recording equipment is required. Therefore, the use of predictive models based on animal information provides a more accessible alternative. Traditionally, Statistical (SA) methods have been employed in the prediction of DC CH_4 emissions. However due to the smart farming revolution, the scale and variety of complex animal information now available for the prediction of DC CH_4 emissions has grown exponentially, and within them are likely to exist non-linear relationships which these traditional SA models may struggle to capture. Therefore, this research aims to explore if Machine Learning (ML) models are a viable alternative for the prediction of DC CH_4 emissions, as they can handle and extract these inevitable non-linear relationships present within today's large, heterogeneous datasets. In this research, we compared a traditional SA method, a Linear Mixed Effects (ME) model, with an original ML method, a Random Forest (RF) model, as well as a novel SA/ML hybrid method, a Mixed Effects Random Forest (MERF) model, in the prediction of CH_4 emissions (CH_4 g/d) produced by DC across 32 experiments. The ML RF model was able to challenge the traditional SA ME model in the prediction of DC CH_4 emissions, achieving a Root Mean Square Prediction Error (RMSPE) and Concordance Correlation Coefficient (CCC) of 52.73 CH_4 g/d and 0.70 respectively, compared to a ME model's 53.90 CH_4 g/d and 0.71. When both the ME and RF models were combined within the novel SA/ML hybrid MERF model, a lower RMSPE and higher CCC were achieved than by each of its composite parts in isolation, 51.87 CH_4 g/d and 0.73 respectively. These results demonstrate the potential of ML in the prediction of DC CH_4 emissions, particularly when hybridised alongside traditional SA methods.

Keywords—Dairy Cattle, Methane, Linear Mixed Effects, Machine Learning, Random Forest, Mixed Effects Random Forest

I. INTRODUCTION

About 30% of the current rise in global surface temperature is due to Methane (CH_4) [1], a potent Green House Gas (GHG) with 84 times the Global Warming Potential (GWP) of Carbon Dioxide (CO_2) over a 20-year period [2]. However, approximately 60% of total global CH_4 emissions are man-made [3]. Whilst much greater than CO_2 in terms of intensity, CH_4 actually has a much shorter atmospheric lifespan, remaining for only 12 years compared to CO_2's centuries [1]. Agriculture is in fact the single largest source of anthropogenic CH_4 emissions worldwide, surpassing even Fossil Fuels, and is responsible for 40% of man-made CH_4 emissions [3]. The overwhelming majority of CH_4 emissions from the Agricultural sector, some 80%, are generated through the unique digestive process which takes place in ruminant Livestock known as Enteric Fermentation (EF) [3]. Of all the ruminant Livestock which produce CH_4 emissions through EF, Cattle alone are responsible for more than all other ruminant Livestock types combined, responsible for 77% of this majority [4]. Due to its intensity, short stay, and largely anthropogenic nature, CH_4 has become a very attractive GHG to target for mitigation, as it can enable a reduction in the Earth's global surface temperature almost within the same decade, allowing the international community to fulfil its collective environmental agreements, whilst still facilitating the growing population.

Whilst there are a myriad of DC CH_4 mitigation techniques, to review the effectiveness of each strategy, one must be able to efficiently record large swathes of DC CH_4 emissions after implementation of the chosen strategy, for assessment. However, the accessibility of DC CH_4 emission recording technologies is a serious barrier to this essential stage of the process. The necessary equipment is expensive and requires a high degree of expertise for implementation, heavily restricting participation.

Therefore, instead of relying on expensive DC CH_4 emission recording technologies, the ability to accurately predict DC CH_4 emissions based on pre-recorded traits using statistical equations has been tirelessly explored, in order to provide perhaps the most efficient methodology for recording DC CH_4 emissions [5]. Rather than through direct measurement, these models predict DC CH_4 emissions based on animal information.

The diversity and depth of biological, environmental, and genetic information which can be used to model the EF process of the animal under representation, hide within their structures and relationships that traditional Statistical (SA) approaches used in the prediction of DC CH_4 emissions may struggle to handle or extract. The sophistication of Machine Learning (ML) on the other hand, has the ability to facilitate these diverse data types and identify the underlying structures and complex relationships between them, which perhaps holds the key to overcoming the limitations currently faced,

and surpassing the accuracies currently achieved, in the prediction of DC CH_4 emissions.

Therefore, in this research, we compared the performance of a traditional SA method, a Linear Mixed Effects (ME) model, with an original ML method, a Random Forest (RF) model, as well as a proposed novel SA/ML hybrid method, a Mixed Effects Random Forest (MERF) model, in the prediction of CH_4 emissions produced by DC across 32 experiments.

This research hopes to highlight the preference of ML in the prediction of DC CH_4 emissions, providing both a more accessible and more accurate alternative to traditional methods. Through these benefits, ML can simplify the validation of multiple CH_4 mitigation strategies and play a vital role in the reduction of Agricultural GHG emissions.

II. Related work

The application of ML in the prediction of DC CH_4 emissions remains extremely rare, although a handful of preliminary investigations have indeed shown the promising potential that ML approaches could bring to the challenge. One study which compared the ability of an ML RF model against a traditional Multiple Linear Regression (MLR) model in the prediction of DC CH_4 emissions, demonstrated the superiority of ML, with the RF model able to consistently outperform the SA MLR model across the multitude of datasets tested [6]. The advantage of ML was noticed particularly in the smaller datasets tested, where the RF model was able to attain a Pearson Correlation Coefficient (r, described below) of 32% in a 3k dataset, while the MLR model was only able to muster an r of 12%. Interestingly, when the MLR model was trained on a much greater dataset, with 41k records, it still could not manage to beat the predictive performance of the RF model trained on the original 3k dataset, achieving an r of 19%, the RF upon this set achieving 71% [6]. The flexibility of the underlying ensemble algorithm that the RF employs is perhaps the reason why this ML model is able to outperform its SA counterpart. Utilising random subsets of predictor variables on bootstrapped samples of the data allows the RF to get a much more intimate understanding of the heterogeneous structures within the dataset, of which it takes an average, laying the foundation for better predictive performance compared to the more rudimentary MLR approach.

As dry matter intake (DMI) increases, enteric CH_4 emission as a proportion of DMI declines [5, 7]. This is due to the increase in rumen passage rate, thus leaving less time for EF of consumed feed in the rumen. This phenomenon strongly encourages traditional SA models to convert from their predominantly linear structure to the non-linear, as through this they will be better able to capture the diminishing increase in enteric CH_4 emissions as feeding level increases.

However, this growing realisation of non-linear relationships between feed intake and CH_4 emissions, which traditional SA models are reluctant to convert to, may be of more value to the ML approach. The diversity and depth of animal information which can be used to model the fermentation biochemistry of the animal under representation, hide within their structures and relationships that traditional linear SA approaches may struggle to handle or extract. With the vast heterogenous datasets being accumulated through the high throughput recording of animal information, as a result of the smart farming revolution [6, 10], the diverse structures and relationships that will inevitably be introduced as a result will challenge a large majority of the traditional SA modelling approaches used in the prediction of DC CH_4 emissions. Yet the equally diverse range of ML algorithms available provides the necessary flexibility required to handle and account for these intricate constraints, facilitating cross talk amongst them, and identifying relationships between them, improving the potential of DC CH_4 emission prediction.

That is not to say that ML models do not also bring with them their own set of challenges to DC CH_4 emission prediction. Another study which tested the ability of an ML Neural Network (NN) model against a SA Partial Least Squares (PLS) model in the prediction of DC CH_4 emissions produced some contrasting results. Taking advantage of the NN's hidden layer activation functions which can accommodate non-linear relationships, the study hypothesised that an NN could better capture the complex relationships of the variables within their dataset, which were subject to causality, non-linearity, or both [11]. As a result, the non-linear variant of the NN was able to outperform the PLS model in the majority of predictor sets tested in the study. Yet, when a Milk Fatty Acid (MFA) predictor was added to the test set, the PLS model overtook the non-linear NN model in terms of predictive performance [11]. The non-linear NN model outperformed the PLS model using the same predictor set based on training data, yet failed to generalise to the test data as reliably as the PLS model [11]. This trend led the authors to believe that the reason for the stall in performance of the non-linear NN model using the MFA predictor set on test data was due to overfitting of the model upon the training set, evidenced by its superior performance in the earlier predictor sets as well as training data of the later predictor sets [11].

Despite being an extremely advanced ML method, deep learning models such as NNs are susceptible to overfitting, especially when the number of neurons in the hidden layer exceeds the number of features supplied to the model [11], impeding predictive performance on new data. Therefore, to avoid such drawbacks, NN models require extensive tweaking before finding the most favourable parameters that can effectively model the data. Just because a model implements ML, this is no guarantee that the model itself will be able to enjoy the benefits of the underlying algorithm. It must be specifically tailored to the context in which it is being applied. Dutifully, one must respect the underlying nuances of the algorithm, in order to get the most out of its potential, when applying it to a predictive problem.

III. Methodology

A. Dataset Understudy

The dataset used in this research was made up of a combination of 32 separate DC CH_4 emission experiments conducted at the Agri Food and Biosciences Institute (AFBI) in Northern Ireland. Each experiment explored a different strategy in the manipulation of DC CH_4 emissions: whether via the effect of experimental dietary treatments, the impact of a selective breed, or the influence of a certain lactation stage, and each varied in scale and duration. A variety of DC breeds were used in each experiment, including Holstein,

Jersey and Norwegian. Respiration Calorimeter Chambers were used in each experiment to record the emission, digestion, and metabolism information of each animal. The physical attributes of each animal, as well as their production level during the experiment, were also available. This resulted in a dataset of 934 observations with 13 available features, made up of 32 experiments and 322 cows (Table I.).

TABLE I. DESCRIPTIVE STATISTICS OF SIGNIFICANT FEATURES

Features	Range		
	Min	*Mean*	*Max*
ECMY (kg/d)	0.90	22.67	45.60
LWT (kg)	379.00	552.50	756.50
Conc/DMI (kg/d)	0	8.15	20.2
CH_4 (g/d)	138.00	375.90	681.30

ECMY = Energy Corrected Milk Yield, LWT = Live Weight, Conc/DMI = Concentrate as a Proportion of Dry Matter Intake, CH_4 g/d = CH_4 Production. Descriptive statistics of full dataset available at: https://computing.ulster.ac.uk/ZhengLab/Ross/

B. Data Preparation

As the interest of this research was in the prediction of DC CH_4 emissions, the feature CH_4 Production (CH_4 g/d) was selected as the target response variable for prediction, and its distribution was visualised for confirmation of normality (Fig. 1).

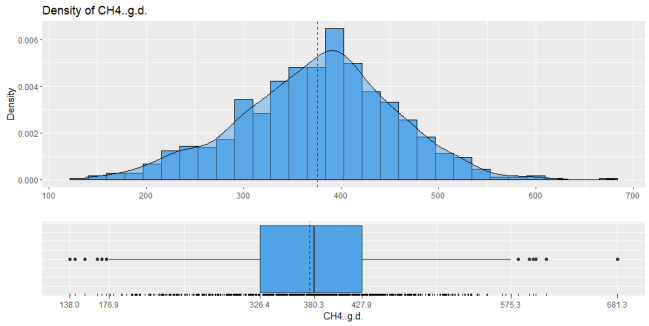

Fig. 1 - Distribution of CH_4 Production (CH_4 g/d)

C. Feature Selection

Dimensionality Reduction

To avoid the curse of dimensionality during model development, a stringent feature selection process was carried out. To initially refine the feature set, a correlation matrix was constructed between each feature and CH_4 g/d using r. R represents the strength and direction of a linear relationship between two features using their covariance and standard deviation (1).

$$r = \frac{n \times (\Sigma(X,Y) - (\Sigma X) \times (\Sigma Y))}{\sqrt{(n \times \Sigma(X^2) - \Sigma(X)^2) \times (n \times (\Sigma(Y^2) - \Sigma(Y)^2)}} \quad (1)$$

Where r is the Pearson Correlation Coefficient, and X and Y are both feature vectors of the same length between which the relationship is sought. The numerator calculates the covariance between X and Y, while the denominator multiplies both of their standard deviations [25]. Ranked between -1 and 1, only those features with a r of over +- 0.25 with CH_4 g/d would be highlighted within the matrix (Fig. 2).

Multicollinearity Handling

To ensure that the features highlighted by the initial correlation matrix were not highly correlated with CH_4 g/d as well as with each other, and thus introduce multicollinearity into our predictive models, this correlation process was hence repeated, this time between the features found to be highly correlated with CH_4 g/d in the initial correlation matrix. Only those features with a r over +- 0.90 between each other would be highlighted within the multicollinearity correlation matrix.

After the multicollinear correlation matrix had been generated, each pair of correlated features was compared with CH_4 g/d, and of the two multicollinear features, only the one more highly correlated with CH_4 g/d, was kept, dropping its correlated feature partner and thus removing the possibility of multicollinearity in future models, as only those features which were highly correlated with CH_4 g/d and also independent of any other feature involved, were kept (Fig. 2).

Feature Significance

To finalise the feature selection process, the concluding stage involved calculating the significance of each remaining feature after Dimensionality Reduction and Multicollinearity Handling through Hypothesis Testing. Whilst a Linear Regression (LR) model would commonly be used for this essential stage of the process; due to the structure of the dataset being used, which was made up of 32 separate Experiments each imposing a direct influence upon CH_4 g/d, the dataset was therefore not independent, violating the primary assumption of the LR algorithm.

Instead, a Linear Mixed Effects (ME) model was chosen, as this could incorporate both fixed and random effects. This allowed the influence of each Experiment and Cow that made up the dataset to be stabilised, and an unbiased view of the influence the independent features of interest had over CH_4 g/d to be obtained. This would result in more precise feature significance calculations. All of the features remaining after the Dimensionality Reduction and Multicollinearity Handling stage were passed into the ME model as fixed effects, with Experiments and Cow being passed in as random effects. The model was trained on the entire dataset.

Only those features which had a p-value of 0.05 or lower within the summary output of the ME model were kept, and these would act as the explanatory features during model development and prediction (Fig. 2).

D. Assessment Pipeline

Using only the features deemed significant during the initial Feature Selection stage, along with the CH_4 g/d response and Experiment and Cow random effects, the dataset was then split into proportional 75%/25% train/test subsets. The models chosen for analysis would be developed using the 75% train set and evaluated using the 25% test set.

The Root Mean Square Prediction Error (RMSPE) and Concordance Correlation Coefficient (CCC) of each model upon the test set would then be calculated for an assessment of their performance. This process would then be repeated ten times, each time refreshing the train/test split and models developed, and an average of each assessment metric over the ten executions was taken (Fig. 3).

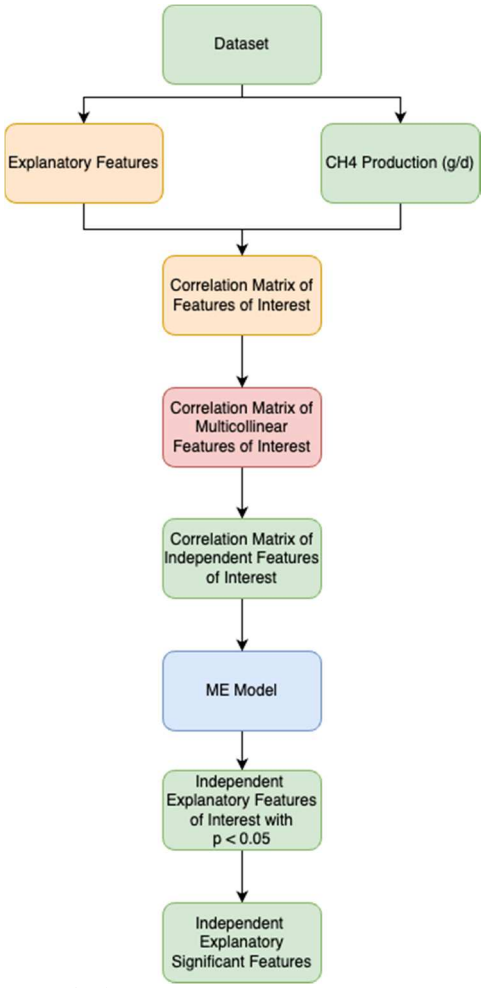

Fig. 2 - Feature selection process

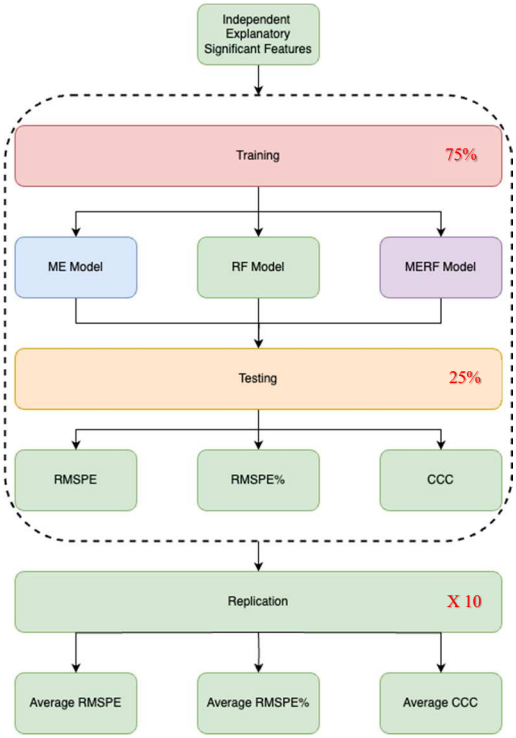

Fig. 3 - Assessment pipeline process

E. Prediction Models

As this research was concerned with the ability of ML to challenge traditional SA methods in the prediction of DC CH_4 emissions, a specific prediction model was selected to represent each approach.

Statistical: Linear Mixed Effects Model

$$y = aX + b_iZ + e \quad (2)$$

Acting as ambassador for the traditional SA approach used in the prediction of DC CH_4 emissions would be a ME model. This would provide the SA benchmark with which the original ML models applied could be compared. The ME model is outlined in (2), where y is the response variable, aX is a set of fixed effects (X) and their coefficients (a), and b_iZ is a set of random effects (Z) with a coefficient (b) for each of their hierarchical subgroups (i) [21].

A ME model initially generates an intercept for the entire sample, much like LR, yet also goes on to generate a unique intercept for each hierarchical subgroup of the random effects. Then, during prediction, the intercept is dynamically updated, starting with the intercept for the entire sample, which is then adjusted accordingly, based on the unique intercept of the hierarchical random effect subgroup the observation under prediction belongs to. This corrects the variation each random effect subgroup imposes upon the fixed effects, and as a result, improves prediction potential.

Machine Learning: Random Forest Model

The prediction of DC CH_4 emissions from the ML perspective would be initially represented by a RF, as this model had been successfully applied within the literature and produced promising results [6], which we wanted to validate and further enhance within this research. A RF comes from the ensemble family of ML algorithms, in which it trains multiple models, in this case, non-linear Regression Trees (RT), on bootstrapped samples of the training data, utilising random subsets of predictor variables. Once all of the RTs have been trained, the RF will then take an average of all of their results. Whilst a RT may be a "weak learner" on its own, when corralled by a RF, it finds its strength in numbers, and an average of all its estimations, makes for a much more accurate prediction, rather than a suggestion when on its own.

A hyper grid was established containing all possible permutations of the RF based on the current dataset structure, and a grid search was then carried out to identify the optimal RF parameters for maximal predictive performance.

Hybrid: Mixed Effects Random Forest

$$y = f(X) + b_iZ + e \quad (3)$$

A SA/ML hybrid method has never before been applied in the prediction of DC CH_4 emissions, leading to the novel implementation of the MERF model used in this research [16, 21]. The MERF model is outlined in (3), where b_iZ represents the same random effect hierarchical subgroup coefficients as in (2), however instead of being added to a set of fixed effect coefficients (aX), they are instead, offset by the output of a general non-linear function, in this case, a RF [21].

The hybrid nature of this model would allow for the random effects of the ME model to be incorporated with the ensemble technique of the RF model, essentially combining the best aspects of both approaches.

This is visually represented in Fig. 4, taking place across two isolated stages, before coming together at the end. In Stage 1, the MERF model trains a RF using the given features, and then uses the trained model to make predictions upon the test set. In Stage 2, the MERF model then trains a ME model, generating a unique intercept for each hierarchical subgroup of its random effects. However, instead of using the linear, PLS algorithm native to the ME model for predictions on the test set, it uses the original, non-linear RF predictions from Stage 1, as an offset. Thus, adding the RF prediction to the uniquely generated intercept of the hierarchical random effect subgroup which the observation under prediction belongs to, allowing it to further correct the original prediction of the RF in Stage 1.

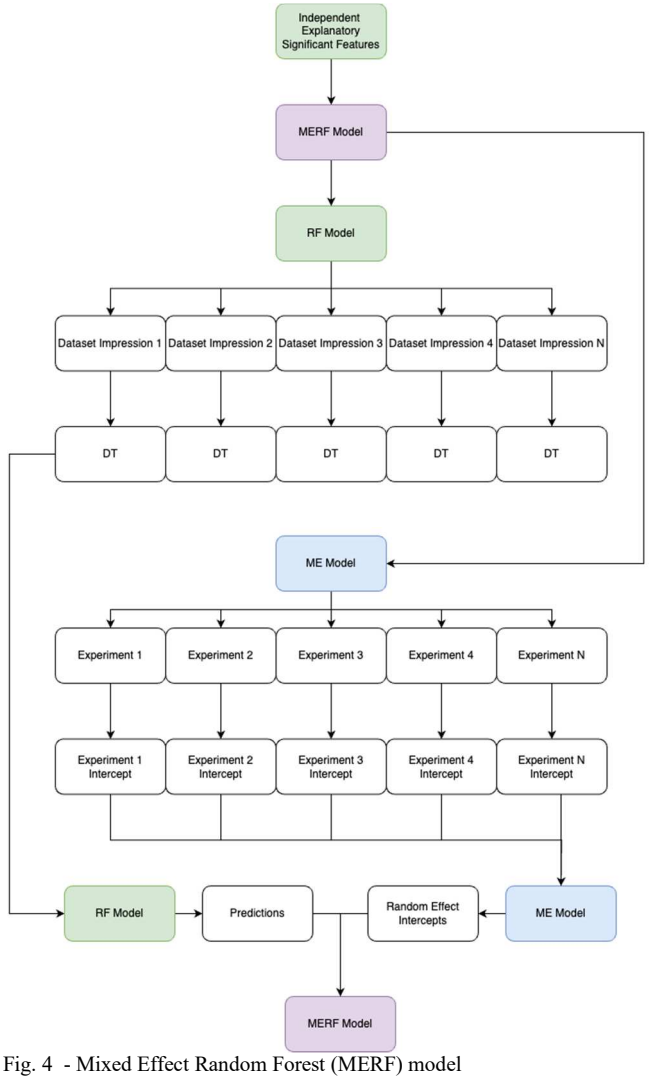

Fig. 4 - Mixed Effect Random Forest (MERF) model

F. Evaluation Metrics

Two key evaluation metrics were used to assess the performance of the models applied in the prediction of CH_4 g/d.

Root Mean Square Prediction Error

The RMSPE is an average of the distance each model's predicted value is away from the actually observed value under prediction. It is expressed in the same scale as the response (CH_4 g/d), however for an additional perspective, this value was also expressed in a percentage form of the response mean (RMSPE%). Lower values here represent a model with higher predictive accuracy. The RMSPE is outlined in (4) where y_i is the actual value of the i^{th} observation, $\overline{y}$; for the predicted, and N represents the total number of observations once the number of parameters (P) have been corrected [22].

$$RMSPE = \sqrt{\frac{\sum(y_i - \overline{y}_i)^2}{N-P}} \quad (4)$$

Concordance Correlation Coefficient

Whilst r is an appropriate metric for evaluating the relationship between two features, to appropriately assess the performance of a prediction model, it is the agreement, not the relationship, between the observed and predicted values, that should be assessed. Therefore, the CCC quantifies the deviation of two vectors of bivariate pairs from a line of perfect concordance (a 45-degree line through 0 on a scatterplot). If both vectors are in complete agreement, their position and scale will be completely identical, and this will allow all values to sit on a perfect line [23].

The CCC is outlined in (5) where y is a vector of observed values, x is a vector of predictions of y, and $cov(x, y)$ is their covariance. The CCC ranges from -1 to 1, with 1 representing a perfect agreement between observation and prediction.

$$CCC = \frac{2cov(x,y)}{(\mu_x - \mu_y)^2 + \sigma_x^2 + \sigma_y^2} \quad (5)$$

G. Implementation

This study was implemented in R. The ME model was implemented using the lme4 package [12]. Due to the implication of the random effects, the lme4 package is reluctant to provide p-values, therefore the lme4 package was extended using lmerTest [13]. lmerTest calculates p-values for the fixed effects using a t-test based on the degrees of freedom approximated by Satterwaite's method [24]. The lmer4 and lmerTest summary was refined and scaled using the jTools package [14].

The RF model was implemented using the ranger package [15], while the MERF model was implemented using the SAE forest package [17].

The dataset splitting process was implemented using the caret package [20].

The MAE and RMSPE metrics were implemented using the Metrics package [18]. The CCC metric was implemented using the DescTools package [19].

IV. Results & Discussion

A. Feature Selection

After the Dimensionality Reduction stage, of the original 13 features in the dataset, 4 were found to have a r of +- 0.25 with CH_4 g/d. Energy Corrected Milk Yield (ECMY kg/d), Milk Yield (MY kg/d), Proportion of Concentrate in the Diet (Conc/DMI kg/d) and Live Weight (LWT) (Fig. 5).

During the Multicollinearity Handling stage, of the 4 features of interest, there was found to be only 1 multicollinear relationship embedded within them, this occurred between the features ECMY kg/d and MY kg/d, which together had a r of 0.93. After the r of each feature in the correlated feature pair was checked against CH_4 g/d individually, ECMY kg/d was found to have a higher r (0.45) with CH_4 g/d, than MY kg/d (0.39) and was therefore kept whilst MY kg/d was dropped. These were added alongside the remaining features from the original 4 of interest which were not correlated with any other features, resulting in 3 features independently correlated with CH_4 g/d in total.

Fig. 5 – Correlation Matrix between Features and CH_4 Production (CH_4 g/d).

To calculate the significance of these 3 features, they were passed into a ME model with the Experiments and Cow features attached as random effects, and CH_4 g/d as the response. The model was trained on the entire dataset. The ME model found all 3 features to be significant in explaining the variation in CH_4 g/d, all with p-values < 0.001. After Dimensionality Reduction, Multicollinearity Handling, and Feature Significance identification, the original feature set was reduced from 13 to 3, which rose to 6 once the CH_4 g/d response, and Experiment and Cow random effects, were reattached.

B. Prediction Performance

After the execution of the Assessment Pipeline using the specific features and models identified, it appears that ML models do indeed have the potential to challenge traditional SA methods used in the prediction of DC CH_4 emissions.

The RF model achieved comparable results with the traditional ME model in both the RMSPE and CCC metrics, in which the RF achieved an average of 52.73 CH_4 g/d and 0.70 respectively, compared to the ME models 53.90 CH_4 g/d and 0.71 (Table II.). Whilst one would have to pick between the higher average CCC provided by the ME model, or the lower average RMSPE provided by the RF model, the novel SA/ML hybrid MERF model applied in this research, allowed for each of its internal model preferences to be fully realised within the same model instead, achieving both a higher CCC than the RF and lower RMSPE than the ME when on their own, 0.73 and 51.87 CH_4 g/d respectively.

The benefit of the MERF model can be more clearly seen within the respective metric scores across the ten tests performed (Fig. 6/7).

In RMSPE (Fig. 6), it appears that the ME model was more precise, having an interquartile range (IQR) of 2.95 CH_4 g/d compared to the RF models 4.76 CH_4 g/d, yet its scores were consistently higher, that same IQR being between 52.32 CH_4 g/d and 55.55 CH_4 g/d compared to the RF models, 49.66 CH_4 g/d to 54.72 CH_4 g/d. However, the hybrid MERF model, which combined both the ME and RF models together, was able to enjoy each of the advantage they provided together, namely the precision of the ME model and the higher accuracy of the RF model, achieving an RMSPE IQR of 50.31 CH_4 g/d to 53.83 CH_4 g/d.

In CCC (Fig. 7), where, inversely to the RMSPE results, the ME model enjoyed a more accurate IQR, between 0.69 and 0.75, compared to the RF models more precise 0.69 to 0.72, the MERF model again was able to enjoy both of these benefits, with a CCC IQR between 0.71 and 0.76.

TABLE II. AVERAGE MODEL ASSESSMENT METRICS AFTER ASSESSMENT PIPELINE

Metrics	**Models**		
	ME	***RF***	***MERF***
RMSPE (CH_4 g/d)	53.90	52.73	**51.87**
RMSPE% (%)	14.24	13.94	**13.94**
CCC	0.71	0.70	**0.73**

RMSPE = Root Mean Squared Prediction Error, RMSPE% = RMSPE expressed as a percentage of the response mean, CCC = Concordance Correlation Coefficient

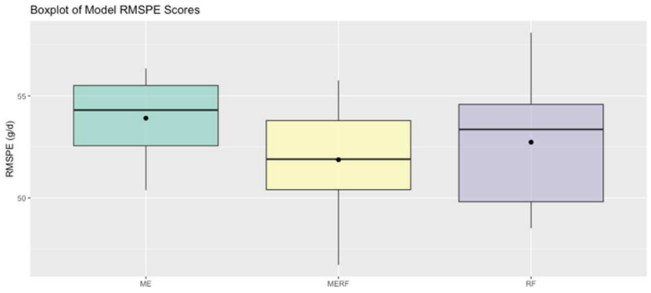

Fig. 6 – Boxplot of RMSPE scores for each model across the ten tests performed (CH_4 g/d).

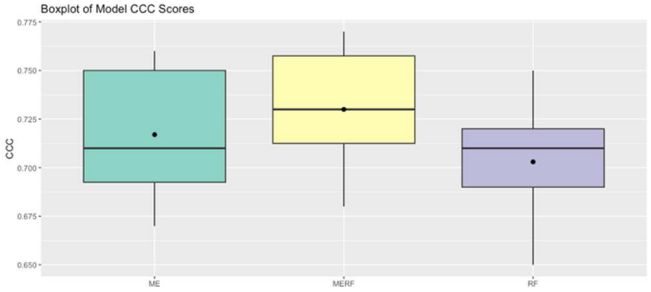

Fig. 7 – Boxplot of CCC scores for each model across the ten tests performed.

This admirable ability of the RF model to challenge the traditional ME model, is likely due to the adaptability of the underlying ensemble algorithm it employs. This algorithm allows for such a comprehensive range of dataset impressions to be made, through bootstrapped samples and feature randomisation, that even without explicitly modelling any random effects, their influence is still encapsulated within the sheer scale of dataset impressions formulated and analysed by the algorithm. So, whilst the RF may not explicitly correct the influence of each random effect, as is the case for with the ME model, its underlying functionality still allows this omission to be captured elsewhere.

However, whilst the RF can certainly still capture a respectable degree of the random effect influence imposed upon the features; through the familarisation of multiple dataset impressions, should they be so extreme, on its own, a RF model will never quite capture the full picture. Here is where the novel SA/ML hybrid MERF model showcases its true potential, as the additional explanation of the random effect influence that the RF model is unable to afford on its own, is selflessly donated by the ME model after the hybrid parley.

No longer stunted by the random effect influence beyond its reach, the non-linear forest of RTs is now able to flourish, and can outperform the more restrictive, linear PLS algorithm native to the ME. It is this complementary exchange that made the novel SA/ML hybrid MERF model the best performing solution in the prediction of DC CH_4 emissions within this research.

V. Conclusion

Due to the complexity and expense of generating large enough DC CH_4 emission datasets, most datasets that define their development are the result of a combination of datasets from multiple experiments, each imposing their own specific influence upon the emissions produced, as well as introducing potential non-linear relationships. Whilst traditional SA methods such as ME models are able to capture the influence of these underlying hierarchies through random effects, their linear nature still struggles to capture the inevitable non-linear relationships inherent within them. ML models on the other hand have shown that they are sophisticated enough to handle both these underlying hierarchies as well as their internal non-linear relationships. Yet it is still a combination of both of these approaches which secures the best performance. In this research, an ML RF model was able to challenge a SA ME model in the prediction of DC CH_4 emissions. But rather than acting as direct competitors, and instead behaving as complimentary teammates, moderated through the SA/ML hybrid model, a MERF, the best aspects of both approaches can be combined, namely the random effects of the ME model and the non-linear ensemble technique of the RF, and as a result, the highest accuracy can be achieved. Whilst an RF, to a certain extent, can begin to sense the influence of any underlying hierarchies via training on multiple dataset impressions, the unique intercepts generated by the ME model still remain the most accurate representation of their individual effect, and if used to temper the predictions made by an RF, will help make them even more refined. The inventive combination of the MERF algorithm within a hybrid model has demonstrated a potentially replicable template, if not a grueling additional step, where these unique hierarchical subgroup intercepts generated by a ME model can be potentially integrated within any ML algorithm. When tested on datasets with more extreme hierarchical influences; as with these large, heterogenous, DC CH_4 emission datasets, which is so often the case; then these hybrid ML model predictions can truly shine. Whilst only preliminary, we believe this study validates the application of ML in DC CH_4 emission prediction, and also illuminates a potential opportunity for their refinement; through a combination of ML techniques and SA methods within a hybrid model.

VI. Future Work

A key area of our future work will be further exploring the potential benefit offered by a combination of ML and SA methods in the prediction of DC CH_4 emissions. It will be interesting to see how predictive accuracy is affected when the unique intercepts generated by a SA ME model, are attached to alternative ML models. The hybrid MERF model studied in this research has allowed for theoretical augmentations of additional popular ML models to be made, such as (ME +) Support Vector Machines, (ME +) Neural Networks, (ME +) Nearest Neighbours and many more. However as seen from the literature [11], just because a model applies ML, this does not mean that a high predictive accuracy is guaranteed, the algorithm itself needs to be conducive with the data available. Due to the low number of features available in the current dataset, deep learning models may unfortunately be out of reach for now, but when more data becomes available, their performance will be of keen interest. We will initially look to apply these models in the prediction of DC CH_4 emissions in their original form, and then compare their performance when the underlying hierarchical subgroup intercepts have been calculated via a ME model. Then, we will have not only a better understanding of the ability of ML in the prediction of DC CH_4 emissions, and potentially, further evidence of their suitability over traditional SA methods, but also, an irrefutable tool for enhancing their performance. Alongside this, alternative feature selection strategies will be tested, and the different feature sets deemed significant by each strategy will be used by each model under investigation. The performances achieved under each feature set identified will then be compared and assessed. Additionally, to cement the position of ML over the traditional SA models used in the prediction of DC CH_4 emissions, non-linear variants of the traditional SA models will also be tested, to ensure that the original ML models being developed, are being compared to the best variations that the traditional SA approach has to offer.

VII. Acknowledgements

This study is a collaboration between Ulster University and the Agri-Food and Biosciences Institute (AFBI) and is funded by the Department of Agriculture, Environment and Rural Affairs (DAERA) through a Postgraduate Studentship.

VIII. References

[1] IEA (2022), Global Methane Tracker 2022, IEA, https://www.iea.org/reports/global-methane-tracker-2022/methane-and-climate-change

[2] European Commission (2022), Methane Emissions, European Commission, https://energy.ec.europa.eu/topics/oil-gas-and-coal/methane-emissions_en.

[3] United Nations Environment Programme and Climate and Clean Air Coalition (2021). Global Methane Assessment: Benefits and Costs of Mitigating Methane Emissions. Nairobi: United Nations Environment Programme.

[4] Gerber, P.J., Steinfeld, H., Henderson, B., Mottet, A., Opio, C., Dijkman, J., Falcucci, A. & Tempio, G. 2013. Tackling climate change through livestock – A global assessment of emissions and mitigation opportunities. Food and Agriculture Organization of the United Nations (FAO), Rome.

[5] Mills, J.A.N., Kebreab, E., Yates, C.M., Crompton, L.A., Cammell, S.B., Dhanoa, M.S., Agnew, R.E. and France, J., 2003. Alternative approaches to predicting methane emissions from dairy cows. Journal

of Animal Science, 81(12), pp.3141-3150, doi: https://doi.org/10.2527/2003.81123141x

[6] Negussie, E., O. González-Recio, M. Battagin, A. Bayat, T. Boland, Y. de Haas, A. Garcia-Rodriguez, P. C. Garnsworthy, N. Gengler, M. Kreuzer, B. Kuhla, J. Lassen, N. Peiren, M. Pszczola, A. Schwarm, H. Soyeurt, A. Vanlierde, T. Yan and F. Biscarini. 2022. Integrating heterogeneous across-country data for proxy-based random forest prediction of enteric methane in dairy cattle. J. Dairy Sci. 105:512_4-5140. doi: https://doi.org/10.3168/jds.2021-20158

[7] Kebreab, E., J. France, B. W. McBride, N. Odongo, A. Bannink, J. A. N. Mills and J. Dijkstra. 2006. Evaluation of Models to Predict Methane Emissions from Enteric Fermentation in North American Dairy Cattle. doi: https://doi.org/10.1079/9781845930059.0299

[8] Huhtanen, P., Ramin, M. and Udén, P., 2015. Nordic dairy cow model Karoline in predicting methane emissions: 1. Model description and sensitivity analysis. Livestock Science, 178, pp.71-80, doi: 10.1016/j.livsci.2015.05.009

[9] Ramin, M. and Huhtanen, P., 2015. Nordic dairy cow model Karoline in predicting methane emissions: 2. Model evaluation. Livestock Science, 178, pp.81-93, doi: 10.1016/j.livsci.2015.05.008

[10] Okpara, M.O., 2019. Milk Fatty Acids Estimation by Mid-Infrared Spectroscopy as Proxy for Prediction of Methane Emission in Dairy Cows. Russian Agricultural Sciences, 45, pp.386-392, doi: https://doi.org/10.3103/S1068367419040116

[11] Shadpour, S., T. C. S. Chud, D. Hailemariam, G. Plastow, H. R. Oliveira, P. Stothard, J. Lassen, F. Miglior, C. F. Baes, D. Tulpan and F. S. Schenkel. 2022. Predicting methane emission in canadian holstein dairy cattle using milk mid-infrared reflectance spectroscopy and other commonly available predictors via artificial neural networks. J. Dairy Sci. 105:8272-8285. doi: https://doi.org/10.3168/jds.2021-21176

[12] Bates D, Mächler M, Bolker B, Walker S (2015). "Fitting Linear Mixed-Effects Models Using lme4." Journal of Statistical Software, 67(1), 1–48. doi:10.18637/jss.v067.i01.

[13] Kuznetsova A, Brockhoff PB, Christensen RHB (2017). "lmerTest Package: Tests in Linear Mixed Effects Models." Journal of Statistical Software, 82(13), 1–26. doi:10.18637/jss.v082.i13.

[14] Long JA (2022). jtools: Analysis and Presentation of Social Scientific Data. R package version 2.2.0, https://cran.r-project.org/package=jtools.

[15] Wright MN, Ziegler A (2017). "ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R." Journal of Statistical Software, 77(1), 1–17. doi:10.18637/jss.v077.i01.

[16] Ahlem Hajjem, François Bellavance & Denis Larocque (2014) Mixed -effects random forest for clustered data, Journal of Statistical Computation and Simulation, 84:6, 1313-1328, DOI: 10.1080/00949655.2012.741599

[17] Krennpa, (2022), SAEForest, rdrr.io, https://rdrr.io/github/krennpa/SAEforest/f/README.md

[18] Hammer Ben, Frasco Michael (2018), Metrics: Evaluation Metrics for Machine Learning, https://cran.r-project.org/web/packages/Metrics/index.html

[19] Andri Signorell et mult. al. (2017). DescTools: Tools for descriptive statistics. R package version 0.99.23.

[20] Kuhn, Max (2008). "Building Predictive Models in R Using the caret Package." Journal of Statistical Software, 28(5), 1–26. doi:10.18637/jss.v028.i05, https://www.jstatsoft.org/index.php/jss/article/view/v028i05.

[21] S. Dey, "Mixed Effects Random Forests," Towards Data Science, 2022, [Online]. Available: https://towardsdatascience.com/mixed-effects-random-forests-6ecbb85cb177. [Accessed: Oct. 19, 2023].

[22] J. Frost, "Root Mean Square Error", Statistics by Jim, [Online] Available: https://statisticsbyjim.com/regression/root-mean-square-error-rmse/ [Accessed: Oct, 19, 2023]

[23] A. Gossman, "Concordance Correlation Coefficient", 0-fold Cross-Validation, [Online] Available: https://www.alexejgossmann.com/ccc/, [Accessed: Oct, 22, 23]

[24] Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest Package: Tests in Linear Mixed Effects Models. Journal of Statistical Software, 82(13), 1–26. https://doi.org/10.18637/jss.v082.i13

[25] J. Fernando. "The Correlation Coefficient: What It is, What It Tells Investors" (2023), [Online] Available: https://www.investopedia.com/terms/c/correlationcoefficient.asp