Utilising an ensemble machine learning model to predict animal disease based on its clinical signs.

First Author[a], Second Author[b], Third Author[a,b,*]

[a]First affiliation, Address, City and Postcode, Country
[b]Second affiliation, Address, City and Postcode, Country
[c]Third affiliation, Address, City and Postcode, Country

Abstract

Background: An ensemble is a widely used machine learning model for disease prediction that can improve prediction accuracy by utilising various models' capabilities. An ensemble method would involve training different independent models on the same dataset to predict animal disease based on clinical signs. A particular clinical sign may be able to identify by one model but difficult to recognise by another. The predictions of these two models can be combined to create an ensemble model, which may result in more precise predictions overall. The ensemble method also decreases the chance of overfitting when a model is too closely fitted to the training data and performs worse on incoming, unknown data.

Results: We used ensemble machine learning to forecast common diseases in African cattle based on clinical signs. This paper uses the potential ensemble learning methods adaptive boosting (AdaBoost), extreme gradient boosting (XGB Classifier), and bagging to predict animal disease. Compared to employing a single algorithm, the ensemble approach considerably increased the predictions' accuracy by up to 70–80%. The Top-N criteria threshold is also used to identify potential diseases following the anticipated ensemble. Performance criteria, including accuracy, precision, recall, and F-1 score, are used to compare the efficiency of the investigated ensemble machine learning approaches.

Conclusions: The investigation shows that Adaboost and XGB Classifier are highly effective at recognising diseases on our dataset.

Keywords: machine learning; ensemble model; multi-label classification; imbalanced data; disease diagnosis.

1. Introduction

Veterinary diagnostics is essential in discovering the cause of an animal's disease or condition.. Machine learning (ML) is a type of artificial intelligence that involves training models on a dataset to generate predictions or decisions. ML techniques have been implemented in a variety of ways to aid with veterinary diagnosis. Using machine learning, researchers have created a diagnostic model that can predict the prevalence of specific diseases based on clinical signs or laboratory data. [1]. These models can help diagnose conditions like cancer, diabetes, and heart disease in animals.

ML approaches are generally becoming more and more common in veterinary diagnosis to increase the precision and effectiveness of the diagnostic procedure. ML can be used to forecast the result of treatment, assist in the early diagnosis of diseases, and help choose the best course of action.

1.1. Machine learning for disease diagnosis

Currently, machine learning models are utilised to predict diseases in humans, plants, and animals. The supervised learning model is widely applied to diagnosing diseases to aid in preventing illness outbreaks and lowering the cost, time and resources spent on treatment. Veterinary diagnostics are essential to obtaining ML information for predictive modelling. Although veterinarians use animal signs to identify, diseases are critical for creating the predictive model. Moreover, data in nature are commonly imbalanced in terms of the prevalence of diseases. Therefore, data preprocessing and input framework design are crucial to the model, and both have a consequence on how effectively to forecast diseases. Ensemble machine learning is a powerful technique for improving the accuracy of predictions in various applications, including predicting animal diseases. Ensemble methods combine multiple models' predictions to create a final prediction by averaging the predictions, weighting the predictions based on their accuracy, or using a meta-model to make the final prediction. In addition, the capability of diagnosis can play an essential role in addressing the disease issue. The time of diagnosis and accuracy has increased the importance of treatment. This research applies the ensemble machine learning model to predict livestock diseases in Africa using only clinical signs.

1.2. Previous studies

According to Bhargavi K. (2022) [2], a combination of Bayes optimal classifier, bootstrap aggregating (bagging), boosting, Bayesian model averaging, Bayesian model combination, a bucket of models, and stacking are used in ensemble machine learning to predict zoonotic diseases. The ensemble approach significantly improved the accuracy of the predictions compared to using a single algorithm.

1.3. Problem Statement

This study will focus on improving the accuracy of the classification problem on multi-label and multi-class classification and engaging the expert knowledge with a reliable method on the disease diagnosis data.

This study applied machine learning principles to create a program or system to optimise prediction accuracy to achieve this problem. In addition, some techniques related to imbalanced data are considered to use or the current data characteristic. Moreover, the efficiency of diagnosis methods should improve by engaging expert knowledge in the diagnosis model**.**

1.4. Research Aims

The following objectives are identified:

1. To incorporate expert knowledge into the machine learning (ML) model for animal disease prediction by applying the Delphi approach, systematically finding consensus from experts**,** with available diagnosis information.
2. Using animal signs to predict imbalanced disease data with muti-class classification by using an existing machine learning model with some feature selection analysis approach.

3. Increase the effectiveness of the prediction model by applying the Top-N ranking criteria to choose the predicted output.

1.5. Research Questions

The following research questions will be answered:

1. Which machine learning model is likely suitable for the diagnosis task?
2. Which of the variety of classification approaches will most appropriate for multi-class or multi-label disease diagnosis data?
3. What imbalanced data technique can build an effective disease diagnosis model?
4. In pre-processing process, any feature selection or reduction technique is applied before training the model?
5. How to create the reliable system to engage expert knowledge with predicted model?
6. How we manage the different judgment of expert opinion and data collection before and after training model?

1.6. Expected Contributions or Outcomes

The expected contribution from this study will be proposed in the program or system that can help user prevent disease. Moreover, some special cases from the expert judgement will retrain to improve the model performance.

The expected contribution from this study will be to propose a systematic approach that can help users prevent disease. Only signs and veterinarians' disease diagnosis data (the gold standard target) were used to develop an efficient predictive model and design the input data for this study. Additionally, the top-N threshold criteria are used to narrow down potential disease outcomes so that farmers can get early diagnostic information before the veterinarian arrives.

1.7. Background

Ensemble machine learning is a method that combines multiple models to improve the performance of a prediction task. In predicting livestock animal disease, ensemble methods can be used to improve the accuracy and robustness of the predictions made by the model.

The main contribution of ensemble methods in this context is the ability to leverage the strengths of different models. While mitigating the model weaknesses, such as other models may perform well on different subsets of the data, an ensemble can be used to combine the predictions in a way that results in improved performance overall. Moreover, ensemble algorithm can reduce the variance and bias of the predictions made by the model to reduce the impact of any individual model's errors.

Using machine learning to diagnose disease involves training a model to identify patterns in data indicative of a particular disease. In the case of diagnosing diseases in humans, plants, and animals[2-7]. The data to train the model would typically include clinical signs, such as symptoms, lab test results, and imaging data [2-6]. The advantages of using machine learning models for disease diagnosis are not intended to replace human expertise but rather to assist and support the diagnosis process. In addition, they are also not always accurate and should be used in conjunction with other diagnostic tools and clinical expertise. ML models achieved high accuracy and precision in diagnosing animal diseases. However, the limitation of specific animal species and exploring how the results generalise to other species would be interesting.

The advantages of utilising machine learning to predict animal illness include improved Accuracy: Machine learning models can analyse large amounts of data and identify patterns that may be difficult for humans to detect. This ML can lead to more accurate predictions of animal illness, early detection, and reduced costs associated with treatment and care. The value of using machine learning to predict animal illness is that it can help to improve the health and welfare of animals and increase the efficiency of the livestock industry. Additionally, machine learning models can help reduce disease spread and improve food safety by detecting the condition early.

1. The supervised learning model for disease diagnosis

There are several simple supervised machine learning models suitable for binary input data and multi-class classification problems. Supervised learning models that can be used for disease diagnosis:

Logistic Regression: Logistic regression is a linear model commonly used for binary classification tasks, such as diagnosing a disease or predicting the likelihood of a disease. This method is a wildly used and straightforward classification algorithm that works well with binary input data[8]. It can also be extended to handle multi-class classification by using techniques such as one-vs-rest or softmax regression. Logistic regression models can be used to analyse data from patient history, physical examination, lab results, and imaging studies to predict the likelihood of disease.

Support Vector Machine (SVM)[9, 10]: SVM is a supervised learning algorithm that can be used for classification and regression tasks SVMs are an effective algorithm capable of handling both binary and multiclass classification issues[11]. SVM also work by finding the hyperplane that maximizes the margin between the two classes.. It can identify patterns in data indicative of a particular disease by mapping the data to a high-dimensional feature space and finding the optimal boundary between the classes.

Neural Network: Neural networks are a machine learning model that can be used for many tasks, including image recognition and natural language processing. They are beneficial for analysing data from imaging studies, such as X-ray or MRI images, to identify patterns indicative of a particular disease.

Naive Bayes: Naive Bayes is a simple probabilistic classifier which is based on the Bayes theorem with strong (naive) independence assumptions between the features. Although, the model is a probabilistic algorithm that is commonly used for text classification tasks. It assumes that the input features are conditionally independent, simplifying the model and making it computationally efficient. It can be used for binary and multi-class classification tasks and can be trained efficiently, even with large datasets [12].

The decision tree[13, 14]: A decision tree is an algorithm used for classification and regression problems in machine learning. Decision trees are versatile and interpretable models that can handle both binary and multi-class classification problems[15]. They work by recursively splitting the data based on the input features until a decision can be made about the class label. The basic idea behind decision trees is to create a tree-like model of decisions and their associated consequences. Each internal node in the tree represents a “test” on an attribute, each branch represents the outcome of the test, and each leaf node represents a class label. The decision tree algorithm works by recursively partitioning the data into subsets based on the values of the input features. At each internal node of the tree, the algorithm selects the component that provides the most information gain to partition the data. The process is repeated recursively on the subsets until a stopping criterion is met.

K-nearest neighbours (KNN)[16]: KNN is a machine learning algorithm that can be used for classification and regression problems. The basic idea behind KNN is to classify a new data point based on the majority class or average value of its k-nearest neighbours in the feature space. KNN is a type of instance-based learning where the model doesn’t learn a general representation of the data; instead, it stores the training instances and uses them as knowledge for the prediction. A k value that is too small can result in a model that is sensitive to noise in the data, while a k value that is too large can lead to a model that is too smooth and unable to capture the data’s underlying patterns.

Random Forest[16]: Random Forest is an ensemble learning method that uses multiple decision trees to make predictions. Random forests can handle both binary and multi-class classification problems[17]. Model work by constructing multiple decision trees on random subsets of the data and then combining the results to make a prediction. A random forest can handle large amounts of data, deal with high dimensionality, and be used for classification and regression tasks. It can identify patterns indicative of a particular disease and analyse data from patient history, physical examination, lab results, and imaging studies.

Even though, the suitable model for this specific problem will depend on several factors, including the size and complexity of the dataset, the number of input features, and the performance metrics that are optimising. However, this study applied the model that suitable for binary input data and multi-class classification problems in the disease diagnosis binary dataset.

2. Ensemble model

An ensemble model is a machine learning model that combines the predictions of multiple individual models to produce a more accurate and robust final forecast. Ensemble methods can be used for classification and regression problems. The idea behind ensemble models is that by combining the predictions of multiple models, the ensemble can mitigate the limitations of any single model, such as high bias or high variance. Several ensemble techniques, such as bagging, boosting and stacking, can be used to create ensemble models. Ensemble models tend to perform better than individual models, especially when the unique models have similar but not identical errors[18].

Adaptive boosting (Adaboost): Adaboost is one of the most popular boosting algorithms that aim to reduce the bias of a base model by iteratively training multiple instances of that model, each with different weights assigned to the training examples. The basic concept of AdaBoost is to train a sequence of weak models, where each model is qualified to correct the errors of the previous model in the sequence.

Bootstrap Aggregating (Bagging): Bagging is an ensemble method that aims to reduce the variance of a base model by averaging the predictions of multiple instances of that model, each trained on a different random subset of the training data. The basic concept of bagging is to train various models independently on different random subsets in the feature space of the training data and then average (for regression) or majority vote (for classification) the predictions of these models.

Extreme gradient boosting (XGBClassfier or Boost): XGBoost implements a gradient boosting algorithm. This ensemble method aims to reduce the bias of a base model by iteratively training multiple instances of that model, each with different weights assigned to the training examples. It is an optimised version of the Gradient Boosting algorithm and is known for its speed and performance. XGBoost is also particularly effective for binary classification problems but can also be extended to handle multi-class classification [19].

XGBoost is a decision tree-based ensemble method that uses a gradient descent-based optimisation algorithm to minimise the cost function. The basic concept of XGBoost is to build an ensemble of decision trees by adding one tree at a time, where each tree tries to correct the mistakes of the previous trees.

3. Balanced and imbalanced data

The impact of balanced and imbalanced datasets on ensemble machine learning for the diagnosis of livestock animal disease can be significant. A balanced dataset has roughly equal numbers of samples for each class. Therefore, the model has enough data to learn the characteristics of each class and can make accurate predictions. On the other hand, an imbalanced dataset has a disproportionate number of samples for one or more categories. This can lead to a model that is biased towards the majority class and may not be able to predict the minority class accurately.

When using ensemble machine learning to predict livestock animal disease in Africa, if the dataset is balanced and contains an equal number of samples of each illness, the model can learn the characteristics of each disease and make accurate predictions. However, suppose the dataset is imbalanced and contains more samples for one disease than the others. In that case, the model may be biased towards that disease and unable to predict the other diseases accurately.

In practice, imbalanced datasets are common, particularly in medical applications where rare diseases are being diagnosed. Oversampling, undersampling, and synthetic data generation can use to balance the dataset. These techniques can combine with ensemble machine learning to improve the model's performance and accurately predict the minority class.

4. Data preprocessing

Data preprocessing is crucial in machine learning, mainly when working with classification problems. It involves cleaning and preparing the data before it is used to train a model. The goal of data preprocessing is to make the data as clean and consistent as possible so that the model can make accurate predictions. There are several steps involved in data preprocessing for classification problems, including:

Data cleaning: This step involves identifying and removing any missing or incorrect data, such as duplicate records or outliers.

Data transformation: This step involves transforming the data into a more suitable format for the machine learning model. This may include normalising the data, scaling the data, or encoding categorical variables.

Data splitting: This step involves dividing the data into training, validation, and test sets. The training set is used to train the model, the validation set is used to evaluate the model during training, and the test set is used to evaluate the model's performance on new, unseen data.

Data augmentation: This step involves increasing the size of the dataset with new data, which can help to improve the model's performance.

The data will be more consistent from the preprocessing steps, making the model more accurate and less prone to error.

2. Dataset

In this paper all dataset come from the work of Beyene, T.J., et al., (2017)[7]. Using a modified Delphi protocol to identify the most significant diseases and Bayesian algorithms to estimate the related disease probabilities based on a variety of clinical signs being available in Ethiopian cattle, this pilot study investigates the use of a VetAfrica-Ethiopiasmartphone-based application in assisting the diagnosis of cattle diseases.

Table 1. Attributes from the original dataset and their values [7]

Animal Species	Attribute	Possible values
Sheep, goat, and cattle	CaseID	Number
	OriginDBName	Name of expert
	DateOfCaseObserved	Date and time
	Species	Name of animal
	DiseaseChosenByUser	Disease names
	Signs	Sign status: Present (P), absent (A) and unknown (U).

Cattle, goats, and sheep are the livestock species used in this study in South Africa, where agriculture is the core business. Therefore, it is essential to diagnose animal diseases before beginning treatment. The veterinarian's data from the farm relates to the animal and typically includes signs and medical diagnoses, as shown in Table 1. Only animal sign data were included in this study as input for modelling. The disease determined by the diagnostic was used as the gold standard or target value for disease prediction. There are three states for signs in animals employed as input.

Table2. An example of a sheep dataset detail

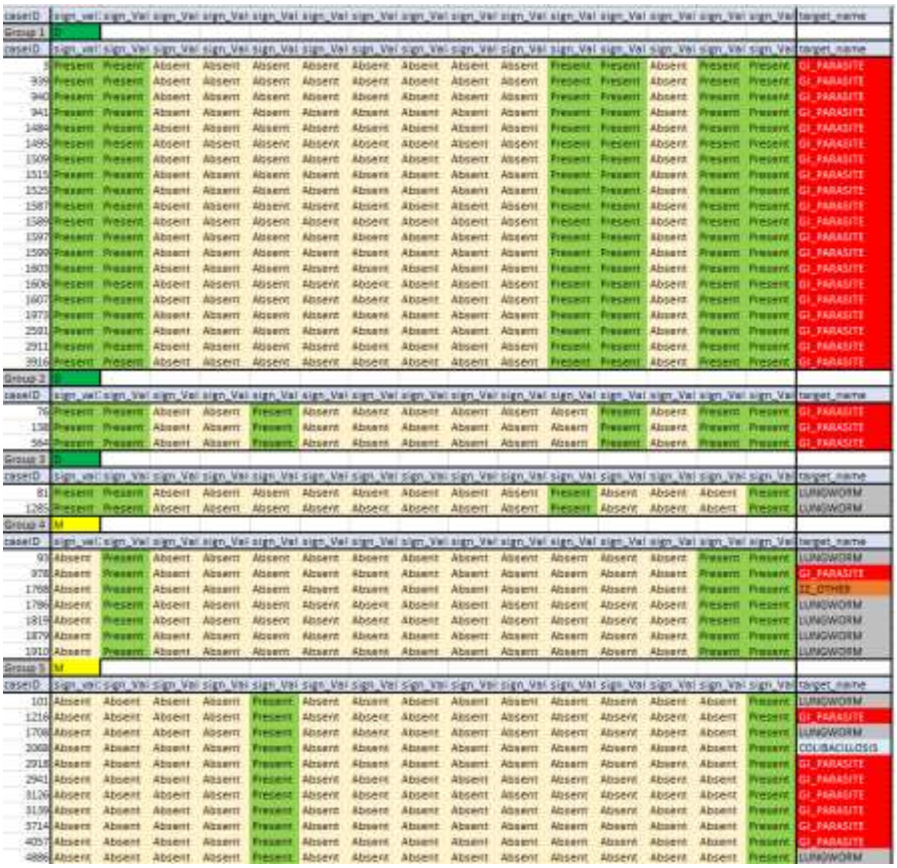
Animal Species	Sign	Expert diagnosis disease
Sheep	1. Anemia / pallor (pale membranes)	1. Coenurosis
	2. Anorexia (loss of appetite)	2. Contagious ecthyma (ORF)
	3. Ataxia / incoordination of movement	3. Cowdriosis
	4. Diarrhea	4. Fasciolosis
	5. Dysentery (blood in faces)	5. Haemonchosis
	6. Dyspnea / coughing (difficulty breathing)	6. Hypocalcemia / Pregnancy tox
	7. Lymph node enlargement	7. Lungworm
	8. Ocular / nasal discharge	8. Mange mite
	9. Pyrexia / fever	9. Nasal bot
	10. Staring coat (standing hair / rough coat)	10. Pasteurellosis
	11. Stunted growth	11. POX
	12. Weakness	12. PPR
	13. Weight loss/emaciation (loss of body condition)	13. Trichostrongiulosis
	14. Constipation	14. Others
	15. Dehydration	
	16. Icterus (yellowing of membranes)	
	17. Submandibular / ventral oedema	
	18. Apathy/depression	

Different signs data were collected for each animal species, as shown in Table 2. There are 18 signs for sheep, and veterinarians have identified 14 diseases (a specific number of diseases, with all other outcomes grouped into an "other disease" category). Although the actual situation by nature, one animal or a case may have more than one disease simultaneously, and the veterinarian will only make one diagnosis in each case. The information used in this study was collected from various subject areas and examples. Some diagnosis sets can be redundant. As a result, this study determined the cases' status in the material displayed in Table 3.

Table3. The designations of case status in a dataset.

Status	Description
Unique (U)	There is only one unique case with only one set of signs and one disease diagnosis.
Duplicated (D)	The cases that have the same set of signs status and target outcome.
Multi-label (M)	The cases have the same signs and more than one single target diseases.

The number and proportion of diseases that occurred in the dataset are different. To depict the prevalence of the disease in nature, there were duplicate cases of the disease (D) with the same set of signs, status, and target outcome. In addition, one animal can have many diseases simultaneously, therefore, there are some cases in which animals experience the same set of input signs, but the disease for which the diagnosis is no single diagnosis. These conditions are referred to as multi-label (M) cases and are shown in Table 3. The unique case (U) refers to some cases with only a set of signs and one disease diagnosis in the dataset. The Unique (U) and Duplicated (D) cases both represent the same cases that has only one single target outcome, but different in the number of cases appear in the dataset. The reason that we separated into two different cases is that the duplicated (D) can represent the prevalence in nature.



caseID	sign_val	sign_Val	sign_Val	sign_Val	sign_Val	sign_Val	sign_Val	sign_Val	sign_Val	sign_Val	sign_Val	sign_Val	sign_Val	sign_Val	sign_Val	target_name
Group 1	D															
caseID	sign_val	sign_Val	sign_Val	sign_Val	sign_Val	sign_Val	sign_Val	sign_Val	sign_Val	sign_Val	sign_Val	sign_Val	sign_Val	sign_Val	sign_Val	target_name
9	Present	Present	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Present	Present	Absent	Present	Present	GI_PARASITE
939	Present	Present	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Present	Present	Absent	Present	Present	GI_PARASITE
940	Present	Present	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Present	Present	Absent	Present	Present	GI_PARASITE
941	Present	Present	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Present	Present	Absent	Present	Present	GI_PARASITE
1484	Present	Present	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Present	Present	Absent	Present	Present	GI_PARASITE
1495	Present	Present	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Present	Present	Absent	Present	Present	GI_PARASITE
1509	Present	Present	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Present	Present	Absent	Present	Present	GI_PARASITE
1515	Present	Present	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Present	Present	Absent	Present	Present	GI_PARASITE
1525	Present	Present	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Present	Present	Absent	Present	Present	GI_PARASITE
1587	Present	Present	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Present	Present	Absent	Present	Present	GI_PARASITE
1589	Present	Present	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Present	Present	Absent	Present	Present	GI_PARASITE
1597	Present	Present	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Present	Present	Absent	Present	Present	GI_PARASITE
1599	Present	Present	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Present	Present	Absent	Present	Present	GI_PARASITE
1603	Present	Present	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Present	Present	Absent	Present	Present	GI_PARASITE
1606	Present	Present	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Present	Present	Absent	Present	Present	GI_PARASITE
1607	Present	Present	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Present	Present	Absent	Present	Present	GI_PARASITE
1973	Present	Present	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Present	Present	Absent	Present	Present	GI_PARASITE
2591	Present	Present	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Present	Present	Absent	Present	Present	GI_PARASITE
2911	Present	Present	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Present	Present	Absent	Present	Present	GI_PARASITE
3916	Present	Present	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Present	Present	Absent	Present	Present	GI_PARASITE
Group 2	D															
caseID	sign_val	sign_Val	sign_Val	sign_Val	sign_Val	sign_Val	sign_Val	sign_Val	sign_Val	sign_Val	sign_Val	sign_Val	sign_Val	sign_Val	sign_Val	target_name
76	Present	Present	Absent	Absent	Present	Absent	Absent	Absent	Absent	Absent	Absent	Present	Absent	Present	Present	GI_PARASITE
138	Present	Present	Absent	Absent	Present	Absent	Absent	Absent	Absent	Absent	Absent	Present	Absent	Present	Present	GI_PARASITE
564	Present	Present	Absent	Absent	Present	Absent	Absent	Absent	Absent	Absent	Absent	Present	Absent	Present	Present	GI_PARASITE
Group 3	D															
caseID	sign_val	sign_Val	sign_Val	sign_Val	sign_Val	sign_Val	sign_Val	sign_Val	sign_Val	sign_Val	sign_Val	sign_Val	sign_Val	sign_Val	sign_Val	target_name
81	Present	Present	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Present	Absent	Absent	Absent	Present	LUNGWORM
1285	Present	Present	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Present	Absent	Absent	Absent	Present	LUNGWORM
Group 4	M															
caseID	sign_val	sign_Val	sign_Val	sign_Val	sign_Val	sign_Val	sign_Val	sign_Val	sign_Val	sign_Val	sign_Val	sign_Val	sign_Val	sign_Val	sign_Val	target_name
93	Absent	Present	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Present	Present	LUNGWORM
978	Absent	Present	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Present	Present	GI_PARASITE
1768	Absent	Present	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Present	Present	IZ_OTHER
1786	Absent	Present	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Present	Present	LUNGWORM
1819	Absent	Present	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Present	Present	LUNGWORM
1879	Absent	Present	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Present	Present	LUNGWORM
1910	Absent	Present	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Present	Present	LUNGWORM
Group 5	M															
caseID	sign_val	sign_Val	sign_Val	sign_Val	sign_Val	sign_Val	sign_Val	sign_Val	sign_Val	sign_Val	sign_Val	sign_Val	sign_Val	sign_Val	sign_Val	target_name
101	Absent	Absent	Absent	Absent	Present	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Present	LUNGWORM
1216	Absent	Absent	Absent	Absent	Present	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Present	GI_PARASITE
1708	Absent	Absent	Absent	Absent	Present	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Present	LUNGWORM
2068	Absent	Absent	Absent	Absent	Present	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Present	COLIBACILLOSIS
2918	Absent	Absent	Absent	Absent	Present	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Present	GI_PARASITE
2941	Absent	Absent	Absent	Absent	Present	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Present	GI_PARASITE
3126	Absent	Absent	Absent	Absent	Present	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Present	GI_PARASITE
3199	Absent	Absent	Absent	Absent	Present	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Present	GI_PARASITE
3714	Absent	Absent	Absent	Absent	Present	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Present	GI_PARASITE
4057	Absent	Absent	Absent	Absent	Present	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Present	GI_PARASITE
4886	Absent	Absent	Absent	Absent	Present	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Present	LUNGWORM

Fig.xx An example of case status of data set

2.1 Data Procedures

Three steps make the dataset processing process: 1) Cleaning the dataset and data preprocessing, which will remove incomplete data, and redundant data.. 2) One-hot encoding techniques will be used to alter information for both input and output in all datasets that we use two structure of input data will be drecribed in only study. 3) Before the algorithm creates a model from the data set, use feature selection to minimise the features of the input. The essential details are related to the dataset as the followings.

1) Data Preprocessing

Since this dataset's information was manually input by the vetenaries , many cases have incomplete data. In the data preprocessing process, we deleted some cases of three species of animal. We kept only the completed case with the same signs and disease limitations. Finally, we removed all those records, such as missing data, signs status being incomplete, all signs' status being absent or unknown value, the target disease may not complete, and the summation likelihood probability may not satisfy. In this work, however, we required only signs and diseases chosen by experts to train the model.

Chi-squared, variance test, and corresponding analysis (CA) are just a few feature selection techniques investigated to rank the most valuable attributes. The goal of feature selection is to select those attributes that are highly dependent on the response. However, we discover that all ranking of signs is the most helpful attribute for target outcome. The number of records in the dataset and the detail are shown in Table 4.

Table 4. The number of cases in each species and the three-case status

Animal	Number of	Number of cases	Number of cases			Number of groups		
Specie	Original cases	after cleaning data	U	D	M	U	D	M
Sheep	1238	832	341	402	89	341	92	29
Goat	789	318	191	94	33	191	31	13
Cattle	3012	1631	532	450	649	532	103	110

Table 4 displays all the dataset's information. There are 1238 case studies in the source data for the sheep dataset. After preprocessing, we exclude any missing or insufficient data, including cases with merely an absent (A) or unknown (U) status. The remaining case after cleaning is 832 cases. The number of cases indicated the proportion of cases in each case status. For duplicate (U) cases, 402 cases in 92 D-groups indicate that sheep have the same signs associated with a single disease. Repeated cases can represent typical behaviour, and the frequency of the disease in nature can serve as a source of potential cases.

2) One-hot encoding technique

Consider our dataset of livestock animals that contains information about their sign and predicting livestock animal disease. Each sign is transformed into two or three binary vectors as *M1* and *M2*. The output signs are the disease that the animal might have. The input signs are converted into a numerical format, which can be fed into a machine learning model. Similarly, the output signs are transformed into the numerical format, which can be used as the target variable for training the model. For example, Anthrax would be represented by a binary vector [1, 0, 0] of three possible diseases, indicating the animal disease as shown in Fig x2.

Before using the one-hot encoding procedure into two structures of M1 and M2 (explained below), the preprocessing data process analyses the duplicated, missing, multi-label situations. We established two structures for training the various data statuses that are

1. *M1:* The present (P), absent (A), and unknown (U) statuses of the sign will be set as 10, 01, and 00, respectively, using two features. Therefore, sheep have 18 indicators that will provide input data for all 36 attributes.
2. *M2:* Similarly, we use three components to present sign status, 100, 010, and 001, for P, A and U status, respectively. In addition, the target outcome is applied to one hot encoding for disease diagnosis.

3) Feature selection

Feature selection is a technique to identify and select a subset of relevant features from a more extensive set of features. The goal of feature selection is to improve a machine learning model's performance by reducing the data’s dimensionality, overfitting, and increasing interpretability. We use feature selection methods with our data set, including Chi2-score and correspondence analysis (CA) (Figure 1).

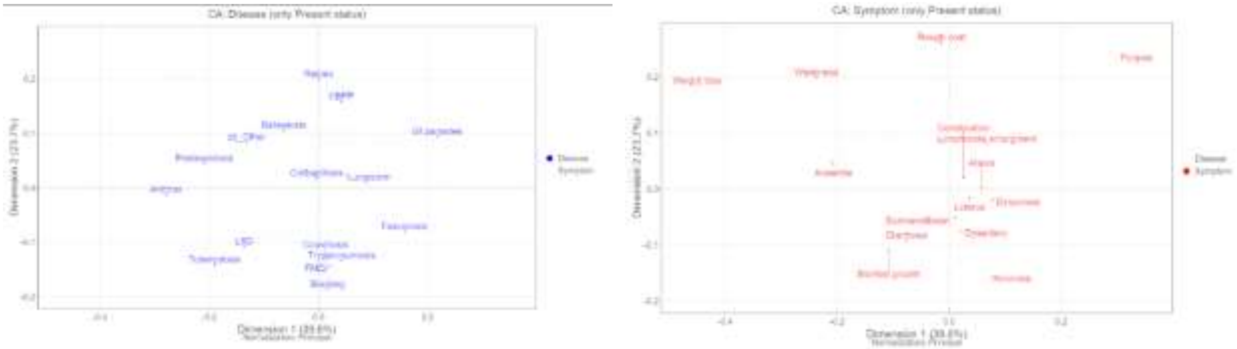


Fig.1 The corresponding analysis (CA) between disease and the present sign status of cattle data set.

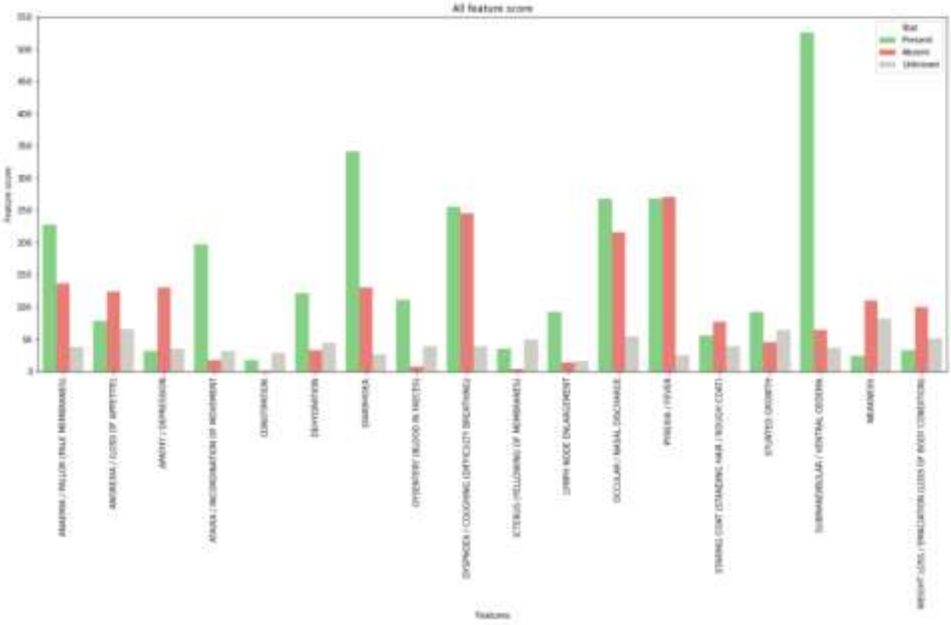


Fig.2 The number of feature and feature score of signs in sheep data set.

In our data set, we applied all about the feature selection method (Figure 2), which are variance, Chi2-schore and corresponding analysis (CA). In this study, Chi-squar is used to calculate the feature score. This method is based on the chi-squared statistical test and it measures the association between each feature and the target variable[20]. In this study we use all attributed as an input is the highest score we got. Therefore, all the input features are fed into the ensemble mode.

2.2 K-Fold Cross-validations

K-fold cross-validation is a widely used technique used to evaluate the performance of a machine learning model to estimate the performance of a model on unseen data. It is a resampling method that allows for the estimation of the model's performance on unseen data. By averaging the performance over multiple partitions of the data, it reduces the variance in the model's performance. It gives a more robust estimate of how the model is likely to perform on new data.

In our study, we applied 5-fold cross-validation by dividing a total of datasets at a ratio of 4:1 to create the training and test sets. The number of cases in the sheep dataset was 832, a relatively small dataset. If the dataset size is small, high variance can cause performance problems for the evaluation of the validation dataset. Therefore, we did not set the validation data ratio for the training set.

2.3 Balanced and imbalanced data

In our imbalanced dataset, we used the Synthetic Minority Over-sampling Technique (SMOTE), a preprocessing method for dealing with an unbalanced class composition.

SMOTE is an oversampling technique used to address class imbalance in a dataset. Class imbalance occurs when one class of data is underrepresented compared to the other class, which can cause problems for machine learning algorithms that assume balanced classes. SMOTE is a method for addressing class imbalance that works by generating synthetic samples of the minority class using the existing samples. The algorithm works by selecting a minority class sample and randomly choosing one or more of its k-nearest neighbors. New synthetic samples are then created by interpolating between the original sample and the selected neighbors.

There are the basic steps of the SMOTE algorithm:

i. Select a minority class sample.
ii. Find k nearest neighbors of the selected sample.
iii. Choose one of the k neighbors randomly.
iv. Generate a new synthetic sample by interpolating between the selected sample and the chosen neighbor.

Repeat steps i-iv until the desired balance between the minority and majority classes is achieved.

The key parameter in the SMOTE algorithm is the value of k, which controls the number of nearest neighbors to use when generating new synthetic samples. Other parameters include the sampling strategy, which controls the ratio of minority to majority class samples in the final dataset, and the random seed, which controls the reproducibility of the results.

Our study uses the Machine learning (ML) model to predict animal diseases which divided into 2 types: Simple and Ensemble models.

2.1. Simple model

There are 3 simple models apply in our dataset which are.

1. Decision Tree (DT)
2. Support Vector Machine (SVM)
3. K-nearest neighbour (KNN)

2.2. Ensemble model

1. Adaptive boosting (Adaboost)
2. Bootstrap Aggregating (Bagging)
3. Extreme gradient boosting (XGBClassfier or Boost)

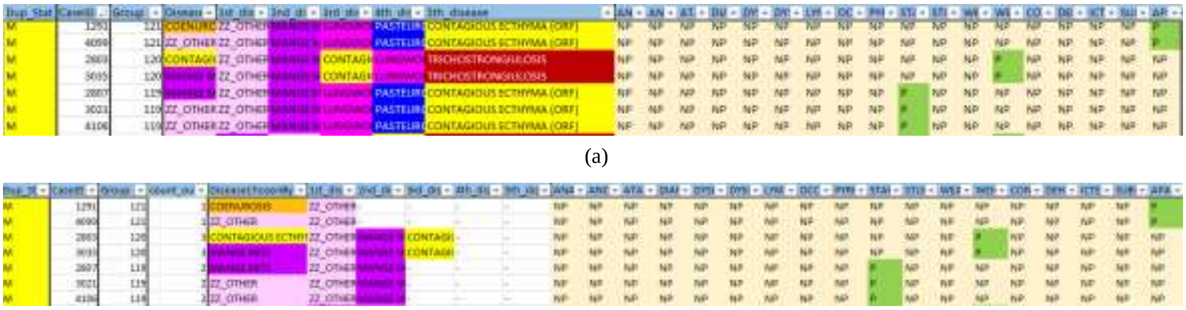
Each model's data was trained using 5-fold cross-validation. We used two different types of data. The original dataset is what we obtained after doing the cleaning operation. The SMOTE dataset is the second type, and the SMOTE imbalanced technique is used with the same species. To compare the variations in these two architectures' effectiveness, each model trains on the identical M1 and M2 input formats.

2.3. Top-N threshold criteria

As a result of the M case status showing a range of expert diagnosis disorders in the same set of signs, our data is also one of the multi-label classification problems. Even though veterinarians only ever identify one disease in a case, an animal may be sick from multiple illnesses simultaneously. In the same circumstance, a different specialist might offer a diagnosis for other diseases. In order to support the output likelihood of the model in our study, we apply the Top-N threshold criteria. The Top-N threshold criteria take into account both the accumulated likelihood in each case and the likelihood of the predicted disease. The Top-N threshold's criteria consist of

1. *The likelihood value of each disease output must be more than or equal to 10% and*
2. *The accumulative likelihood of all disease in the Top-N ranking is less than the threshold criteria*

From these two criteria, we have tried many the accumulation likelihood threshold (ALT) such as 70%, 80% and 90% to compare the number of output diseases and the number of target outcome. Obviously, the highest number of ALT obtains the top performance for every model. Although, the 90% of ALT is always provided more possible output diseases but the number of output diseases from the model are also corresponding to the number of target disease on the training phase.



(a)

(b)

Fig.3 The comparison between the original model output and the applying 90% of ALT on sheep dataset.

From Figure 3, the figure 3 (a) shows the example of output in the multi-label (M) case status that have more than one single target disease. The 1st -5th columns show the predicted output in ranking of likelihood. In Group number 121 mean there are two cases have the same set of signs but different target disease (column Disease target). Figure 3(b) depicted the output that applied the Top-N 90% threshold criteria that the number of predicted output meet the criteria. There are three different diseases in group number 119 and the Top-N technique also provide 3 diseases output.

2.4. Metrics

These metrics are commonly used to evaluate our classification model's performance and compare different models. Depending on the characteristics of the problem, some metrics may be more relevant than others.

Accuracy: It measures the proportion of correct predictions in the total predictions. Accuracy computes the ratio of correctly predicted observation to the total observations. Mathematically, it is represented as (TP + TN) / (TP + TN + FP + FN).

Precision: It measures the proportion of true positive predictions out of all positive predictions made. Precision computes the ratio of correctly predicted positive observations to the total predicted positive observations. Mathematically, it is represented as TP / (TP + FP).

Recall (or Sensitivity or True Positive Rate): It measures the proportion of true positive predictions out of all actual positive observations. A recall is the ratio of correctly predicted positive observations to all observations in the actual class. Mathematically, it is represented as TP / (TP + FN).

F-1 Score: It is the Harmonic mean of precision and recall, and it balances both metrics by considering the trade-off between precision and recall. The range for the F-1 score is [0, 1]. The score shows how accurate the classifier is (how many instances it classifies correctly), how robust it is and does not miss a significant number of instances. The greater the F-1 score, the better our model's performance. Mathematically, it is represented as 2*(Recall * Precision) / (Recall + Precision).

We used precision, recall, F-1 score, recall matrix, and accuracy to assess the performance of the outcomes in our dataset. The precision-recall plot, misclassification matrix, and heatmap are all included in the results visualisation. The precision-recall plot is also appropriate for binary classification issues when our dataset contains imbalanced data [21].

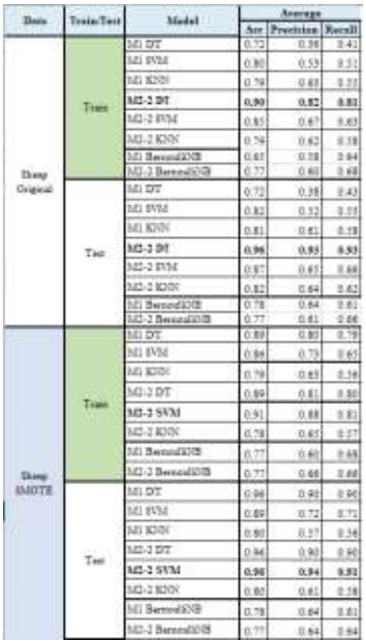
3. Results

This section presents several experimental results as well as a performance evaluation of the suggested method. The outcomes from a simple machine learning model, an ensemble model, and the Top-N threshold criterion will be presented in this study.

3.1. The result from the Simple model

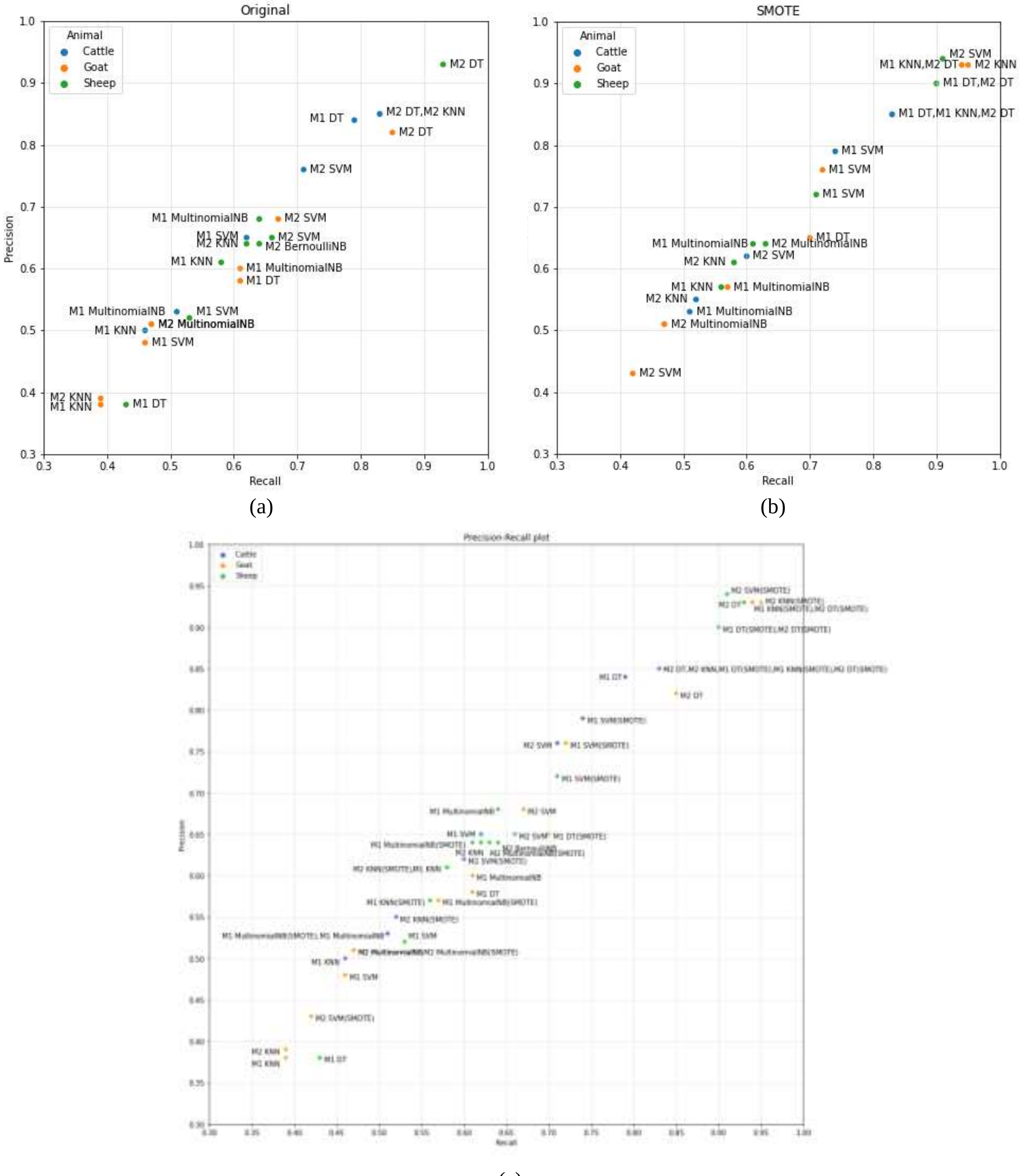
Several ML models have been investigated in this work. We used three datasets of the animal from the VetAfrica-Ethiopia smartphone app for this study. We also looked at ensemble methods as well as conventional approaches like decision trees (DT), support vector machines (SVM), and K-nearest neighbours (KNN). The oversampling method SMOTE is also used to test the model using the two data structures *M1* and *M2* to address the imbalanced problem. Table 5 illustrates the simple models' outcomes.

Table 5 Performance indicator of simple models for predicting sheep diseases in both original and SMOTE datasets, where best results are in bold.



Data	Train/Test	Model	Average		
			Acc	Precision	Recall
Sheep Original	Train	M1 DT	0.72	0.36	0.41
		M1 SVM	0.80	0.53	0.51
		M1 KNN	0.79	0.63	0.53
		M1-2 DT	**0.90**	**0.82**	**0.83**
		M1-2 SVM	0.85	0.67	0.63
		M1-2 KNN	0.79	0.62	0.58
		M1 BernoulliNB	0.65	0.58	0.64
		M1-2 BernoulliNB	0.77	0.60	0.68
	Test	M1 DT	0.72	0.38	0.43
		M1 SVM	0.82	0.52	0.53
		M1 KNN	0.81	0.61	0.58
		M1-2 DT	**0.96**	**0.95**	**0.93**
		M1-2 SVM	0.87	0.65	0.66
		M1-2 KNN	0.82	0.64	0.62
		M1 BernoulliNB	0.78	0.64	0.61
		M1-2 BernoulliNB	0.77	0.61	0.66
Sheep SMOTE	Train	M1 DT	0.89	0.80	0.79
		M1 SVM	0.86	0.73	0.65
		M1 KNN	0.79	0.65	0.56
		M1-2 DT	0.89	0.81	0.80
		M1-2 SVM	0.91	0.88	0.81
		M1-2 KNN	0.78	0.65	0.57
		M1 BernoulliNB	0.77	0.60	0.68
		M1-2 BernoulliNB	0.77	0.66	0.66
	Test	M1 DT	0.96	0.92	0.96
		M1 SVM	0.89	0.72	0.71
		M1 KNN	0.80	0.57	0.56
		M1-2 DT	0.96	0.90	0.90
		M1-2 SVM	**0.96**	**0.94**	**0.92**
		M1-2 KNN	0.80	0.61	0.58
		M1 BernoulliNB	0.78	0.64	0.61
		M1-2 BernoulliNB	0.77	0.64	0.64

Table 5 shows the average accuracy of five folds cross-validation of the sheep dataset. The precision, recall and AUC-ROC values of the simple ML models are evaluated from 5 folds cross-validation. All disease data in all folds have 80% of every disease used in the training data and the remaining 20%, which will be the unseen data used for the testing phase. The highest accuracy scores are from the decision tree (DT) for the original data and SVM for SMOTE dataset.



(a)

(b)

(c)

Fig. 4. Precision versus Recall from table 5 for (a) original dataset and (b) SMOTE dataset and (c) this comparison of both dataset on simple models.

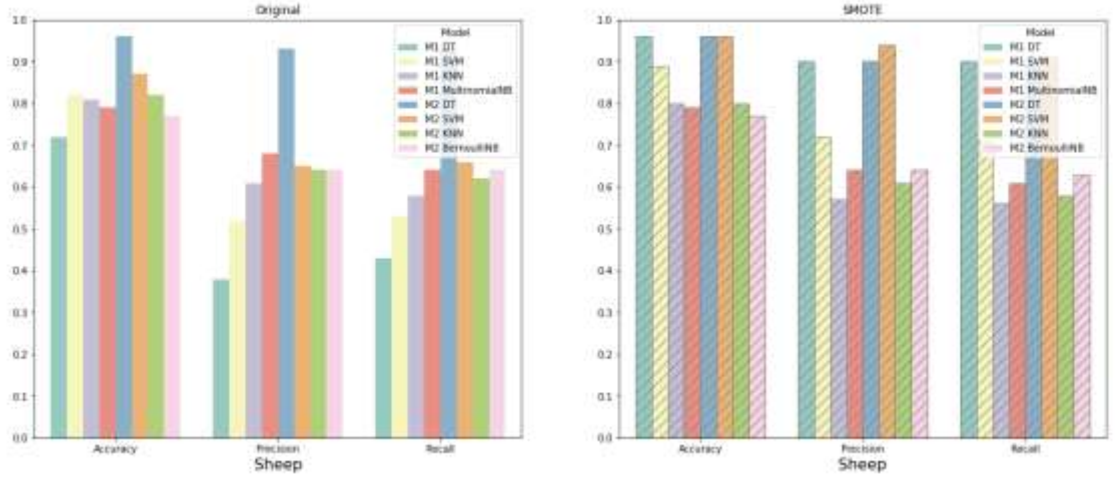
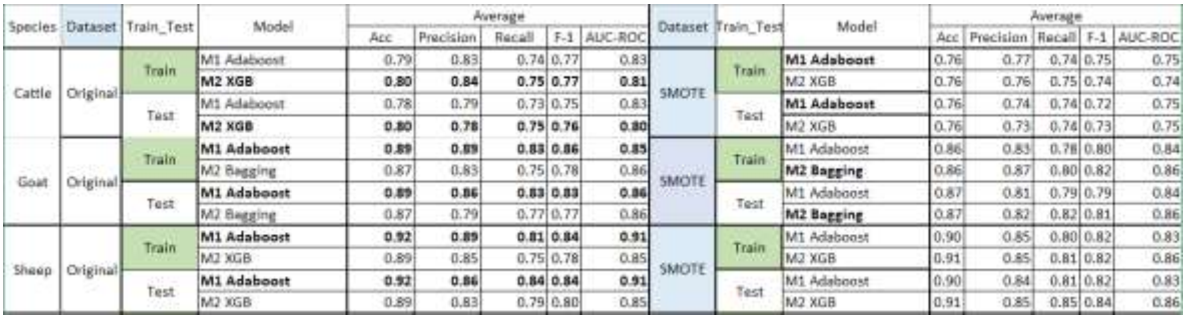


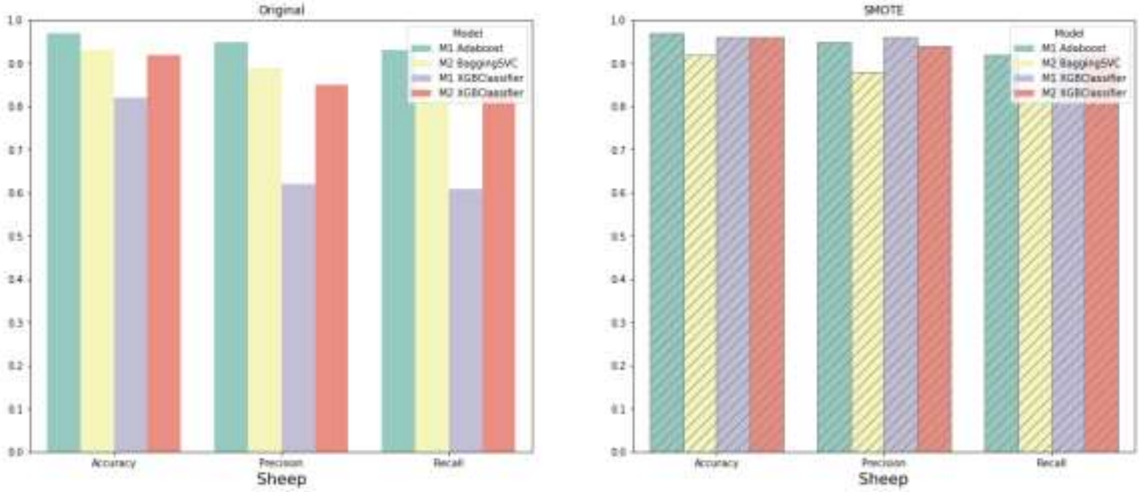
Fig. 5. The comparison between accuracy, precision and recall of sheep dataset with different *M*1 and *M2* input encoding from table 5 for (a) original dataset and (b) SMOTE dataset of simple models.
https://colab.research.google.com/drive/1cOHSTWZFKjlvXRgVusBsylZB-ST6hsA-

3.2. The result from the ensemble model

Table 6: Results of the ensemble model



Species	Dataset	Train_Test	Model	Average					Dataset	Train_Test	Model	Average				
				Acc	Precision	Recall	F-1	AUC-ROC				Acc	Precision	Recall	F-1	AUC-ROC
Cattle	Original	Train	M1 Adaboost	0.79	0.83	0.74	0.77	0.83	SMOTE	Train	**M1 Adaboost**	0.76	0.77	0.74	0.75	0.75
			M2 XGB	**0.80**	**0.84**	**0.75**	**0.77**	**0.81**			M2 XGB	0.76	0.76	0.75	0.74	0.74
		Test	M1 Adaboost	0.78	0.79	0.73	0.75	0.83		Test	**M1 Adaboost**	0.76	0.74	0.74	0.72	0.75
			M2 XGB	**0.80**	**0.78**	**0.75**	**0.76**	**0.80**			M2 XGB	0.76	0.73	0.74	0.73	0.75
Goat	Original	Train	**M1 Adaboost**	**0.89**	**0.89**	**0.83**	**0.86**	**0.85**	SMOTE	Train	M1 Adaboost	0.86	0.83	0.78	0.80	0.84
			M2 Bagging	0.87	0.83	0.75	0.78	0.86			**M2 Bagging**	0.86	0.87	0.80	0.82	0.86
		Test	**M1 Adaboost**	**0.89**	**0.86**	**0.83**	**0.83**	**0.86**		Test	M1 Adaboost	0.87	0.81	0.79	0.79	0.84
			M2 Bagging	0.87	0.79	0.77	0.77	0.86			**M2 Bagging**	0.87	0.82	0.82	0.81	0.86
Sheep	Original	Train	**M1 Adaboost**	**0.92**	**0.89**	**0.81**	**0.84**	**0.91**	SMOTE	Train	M1 Adaboost	0.90	0.85	0.80	0.82	0.83
			M2 XGB	0.89	0.85	0.75	0.78	0.85			M2 XGB	0.91	0.85	0.81	0.82	0.86
		Test	**M1 Adaboost**	**0.92**	**0.86**	**0.84**	**0.84**	**0.91**		Test	M1 Adaboost	0.90	0.84	0.81	0.82	0.83
			M2 XGB	0.89	0.83	0.79	0.80	0.85			M2 XGB	0.91	0.85	0.85	0.84	0.86



https://colab.research.google.com/drive/1fepn9TOpPvfWcN6s2ghYULRLtjoz43Qo?authuser=1

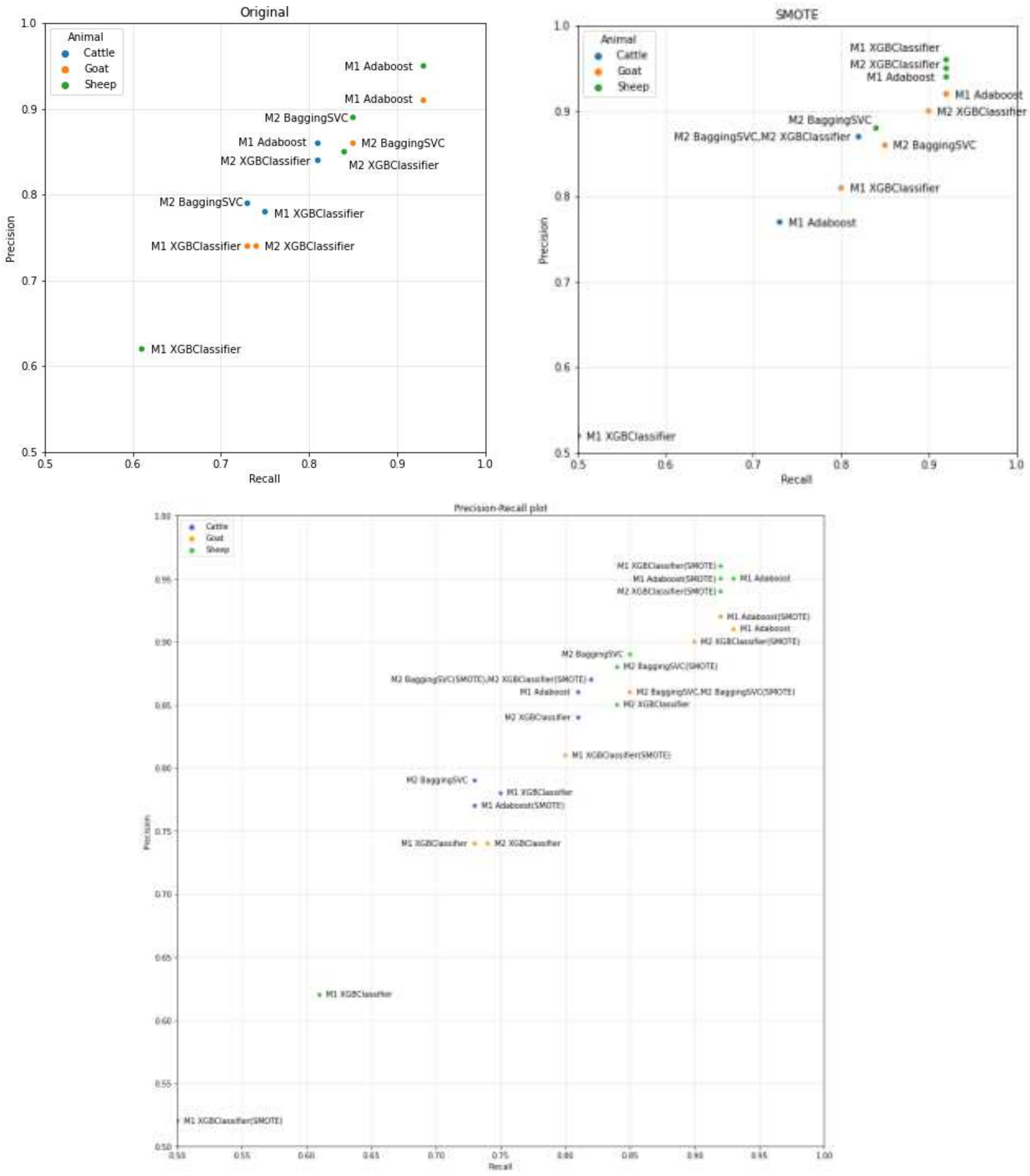

Original
Animal
● Cattle
● Goat
● Sheep
M1 Adaboost
M1 Adaboost
M2 BaggingSVC
M1 Adaboost
M2 BaggingSVC
M2 XGBClassifier
M2 XGBClassifier
M2 BaggingSVC
M1 XGBClassifier
M1 XGBClassifier
M2 XGBClassifier
M1 XGBClassifier
Precision
Recall
1.0
0.9
0.8
0.7
0.6
0.5
0.5
0.6
0.7
0.8
0.9
1.0
SMOTE
Animal
● Cattle
● Goat
● Sheep
M1 XGBClassifier
M2 XGBClassifier
M1 Adaboost
M1 Adaboost
M2 XGBClassifier
M2 BaggingSVC
M2 BaggingSVC,M2 XGBClassifier
M2 BaggingSVC
M1 XGBClassifier
M1 Adaboost
M1 XGBClassifier
Precision
Recall
1.0
0.9
0.8
0.7
0.6
0.5
0.5
0.6
0.7
0.8
0.9
1.0
Precision-Recall plot
● Cattle
● Goat
● Sheep
M1 XGBClassifier(SMOTE)
M1 Adaboost(SMOTE)
M1 Adaboost
M2 XGBClassifier(SMOTE)
M1 Adaboost(SMOTE)
M1 Adaboost
M2 XGBClassifier(SMOTE)
M2 BaggingSVC
M2 BaggingSVC(SMOTE)
M2 BaggingSVC(SMOTE),M2 XGBClassifier(SMOTE)
M1 Adaboost
M2 BaggingSVC,M2 BaggingSVC(SMOTE)
M2 XGBClassifier
M2 XGBClassifier
M1 XGBClassifier(SMOTE)
M2 BaggingSVC
M1 XGBClassifier
M1 Adaboost(SMOTE)
M1 XGBClassifier
M2 XGBClassifier
M1 XGBClassifier
M1 XGBClassifier(SMOTE)
Precision
Recall
1.00
0.95
0.90
0.85
0.80
0.75
0.70
0.65
0.60
0.55
0.50
0.50
0.55
0.60
0.65
0.70
0.75
0.80
0.85
0.90
0.95
1.00

Fig. 6. The Precision versus Recall of the ensemble model

Table 7: Results of Naïve Bayes model

Fig.7 The comparison of Precision versus Recall of all datasets including simple model results on table 5 and ensemble model on table 6.

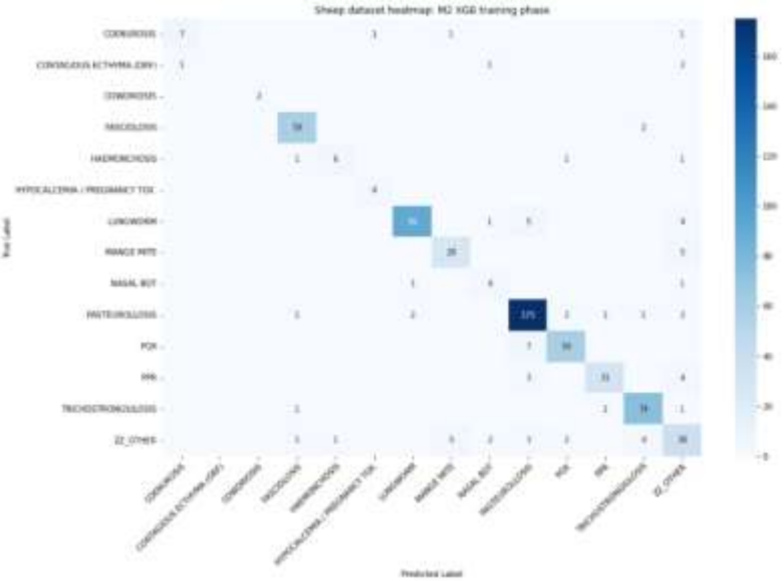


Fig. 8. The confusion ,matrix of the training phase of the sheep dataset on XGB classifier with the M2 encoding structure.

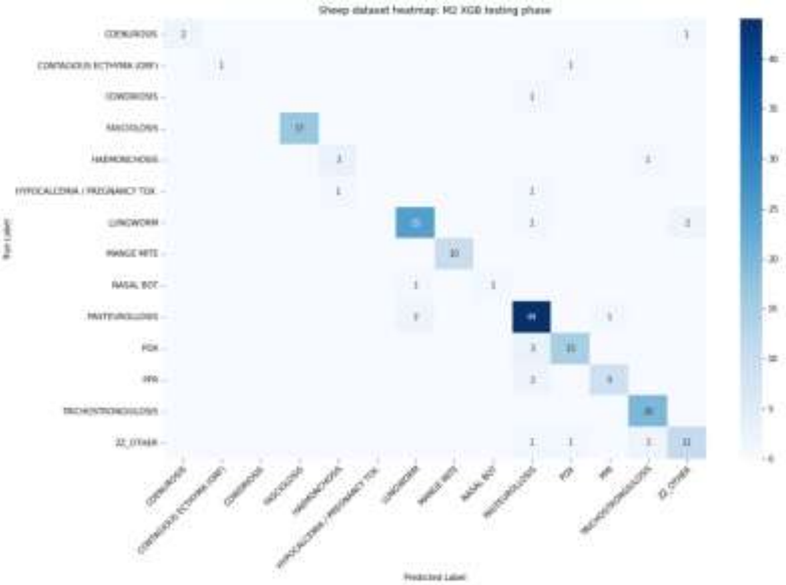


Fig. 9 The heatmap of the testing phase of the sheep dataset on XGB classifier with the M2 encoding structure.

The confusion matrix of sheep dataset are illustrated in Figure 8 and 9 which using XGB classifier on M2 encoding structure. Figure 8 shows that results from the training phase which has 14 signs and 14 target disease classes. The dark blue represents the majority class of sheep disease which is the Pasteurollosis and the lightest blue is the Cowdriosis.

3.3. Top N Threshold criteria

The comparison of different threshold values (70,80, and 90%) shows that the accuracy score is almost similar, and the 90% threshold is achieved as the highest accuracy score on the simple models.

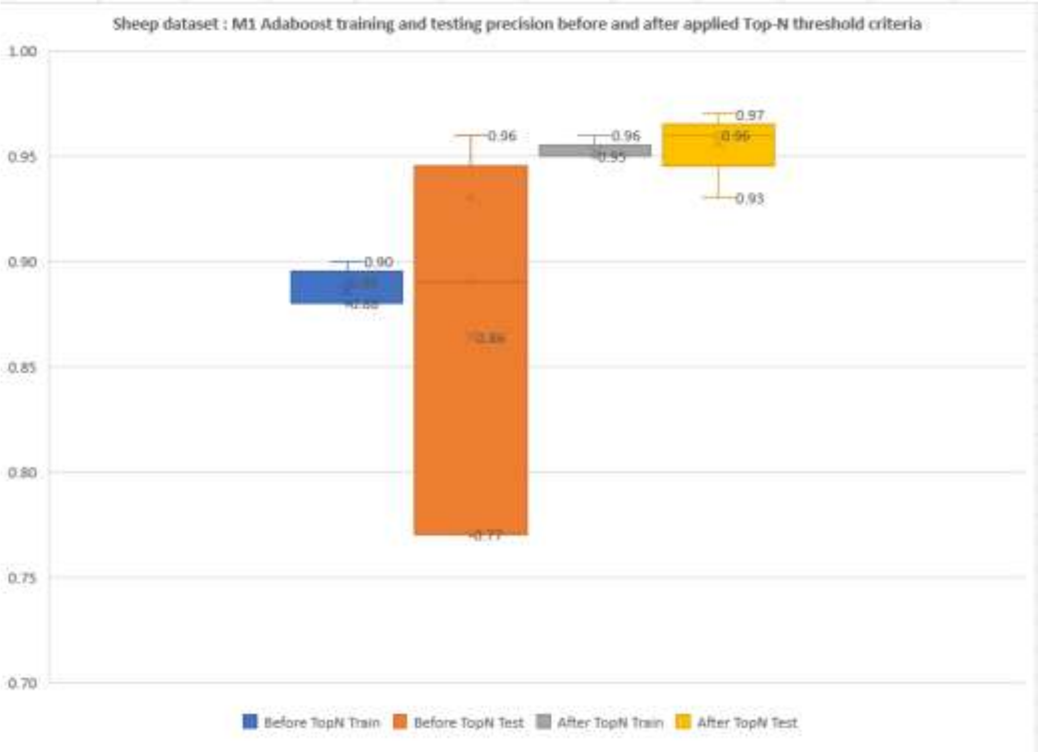


Fig. 13 The average model precision before and after applying the Top N threshold criteria in the training phase.

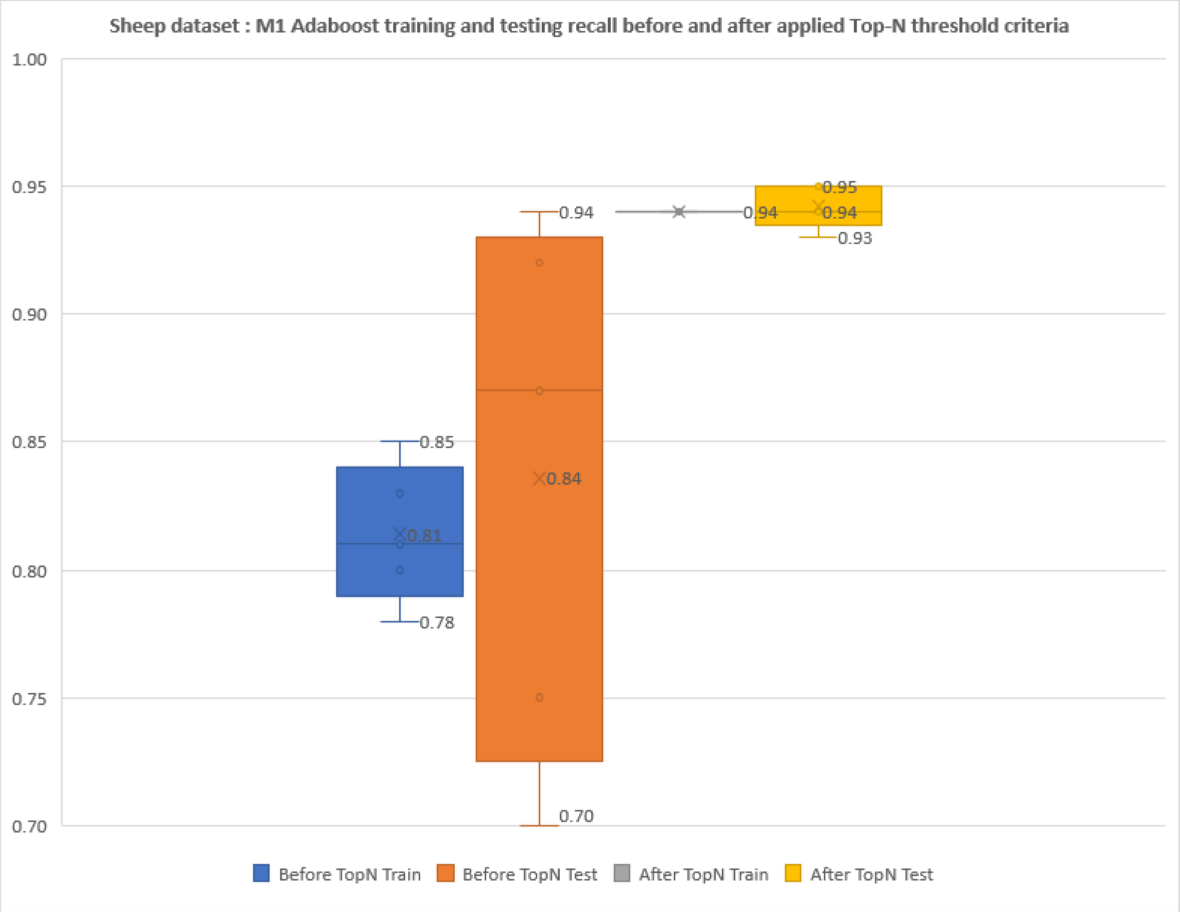


Fig.14 The average model recall before and after applying the Top N threshold criteria in the testing phase.

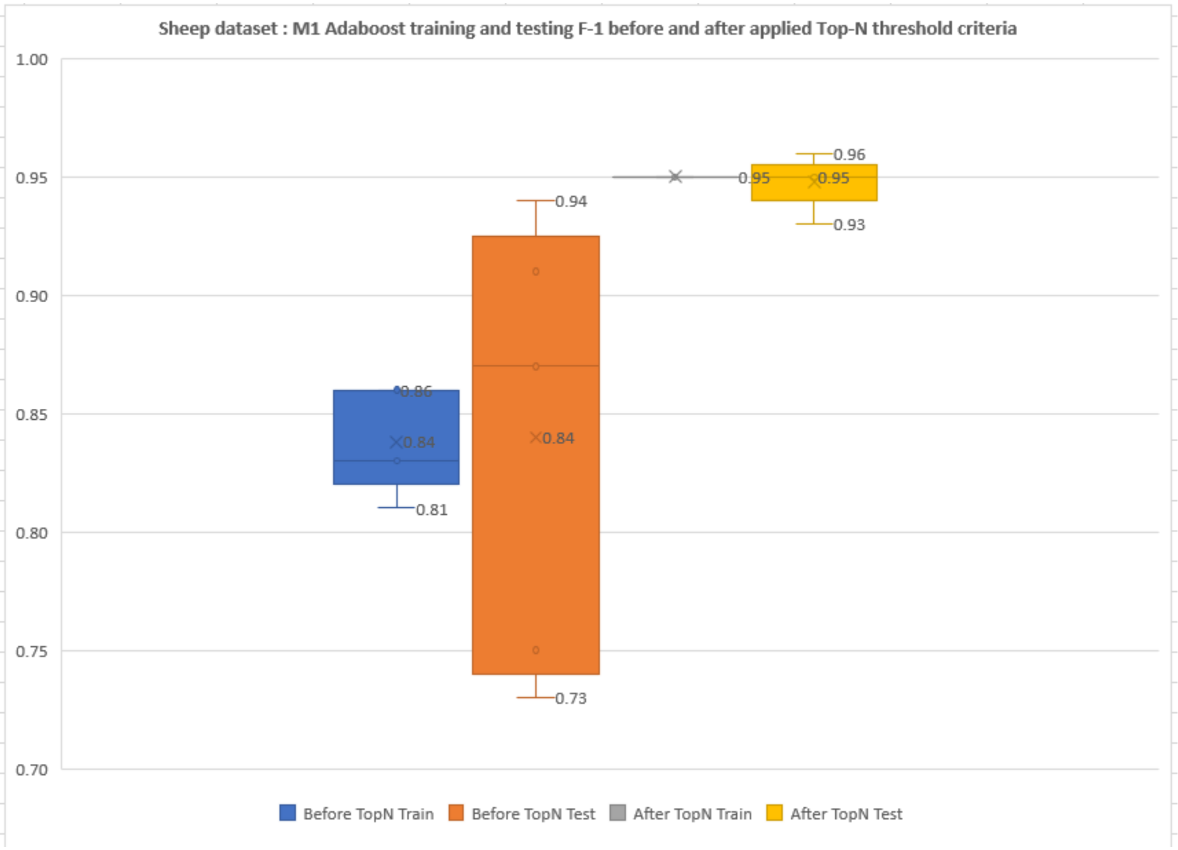


Fig.16 The average model F-1 before and after applying the Top N threshold criteria in the testing phase

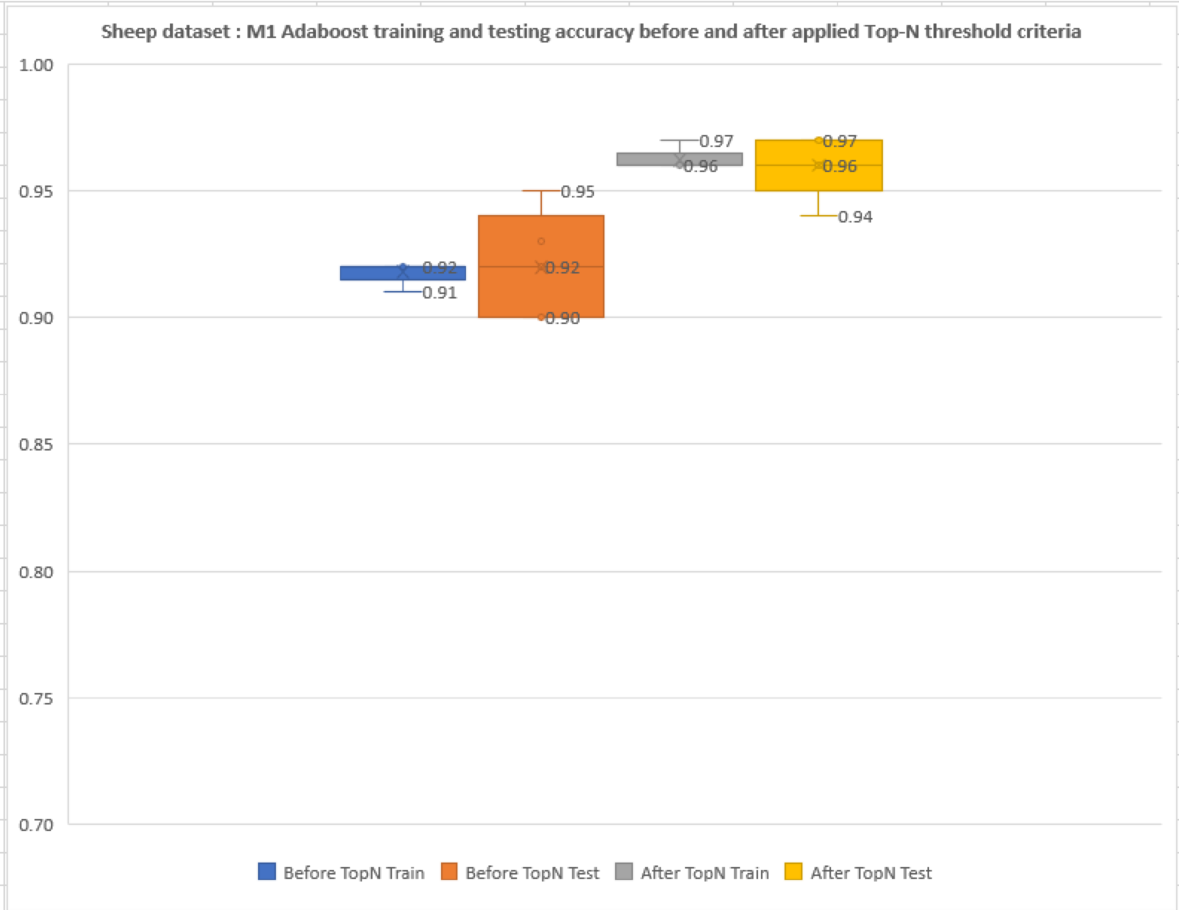
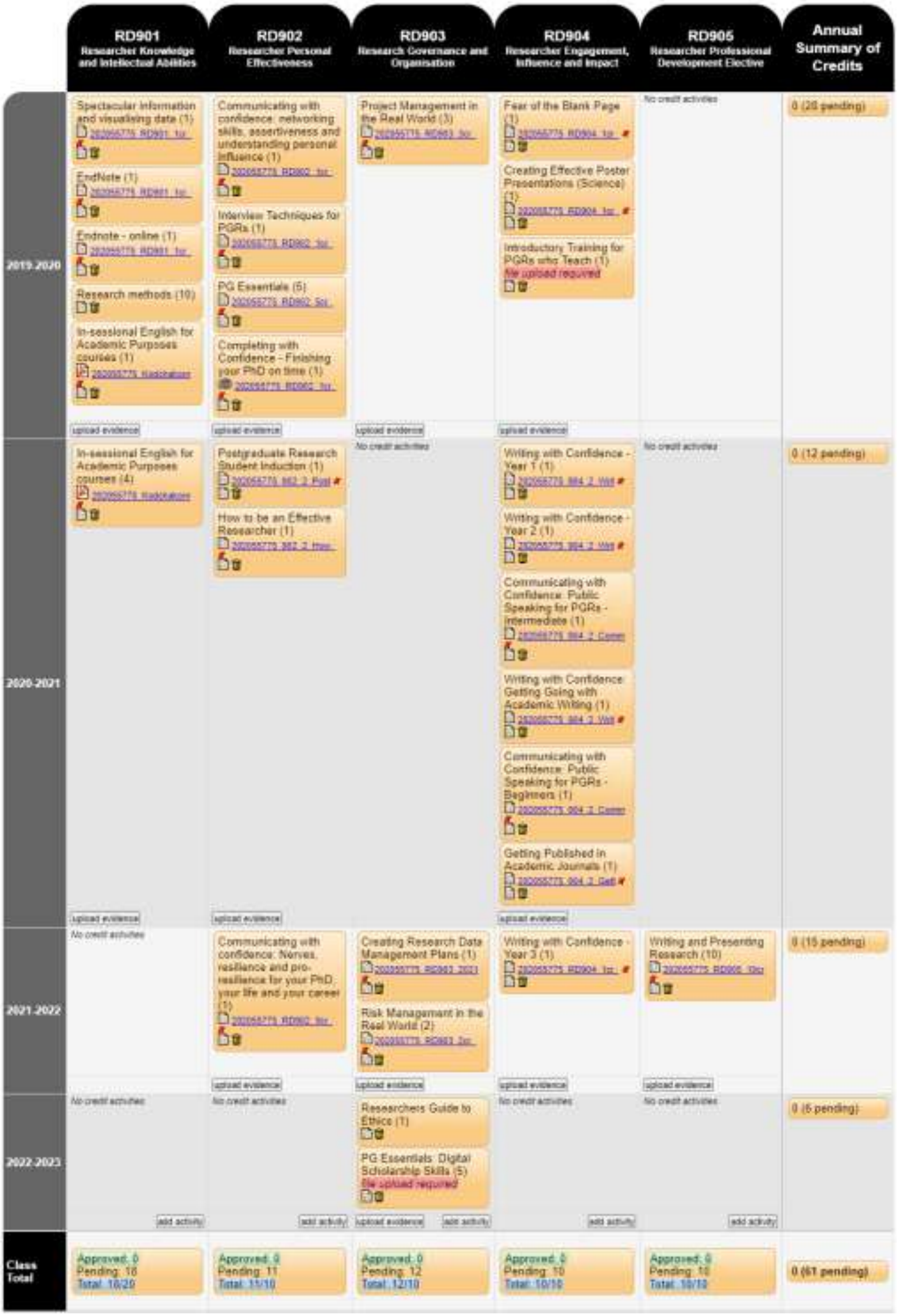


Fig. 16 The average model accuracy before and after applying the Top N threshold criteria in the training phase.

Reference

1. Hassan, F.A., et al., *Machine Learning Based Prediction for Solving Veterinary Data Problems: A Review.* Journal of Advanced Veterinary Research, 2022. **12**(6): p. 798-802.
2. Bhargavi, K., *Zonotic Diseases Detection Using Ensemble Machine Learning Algorithms.* Fundamentals and Methods of Machine and Deep Learning: Algorithms, Tools and Applications, 2022: p. 17-32.
3. Ebrahimi, M., et al., *Comprehensive analysis of machine learning models for prediction of sub-clinical mastitis: Deep Learning and Gradient-Boosted Trees outperform other models.* Computers in biology and medicine, 2019. **114**: p. 103456.
4. Subbulakshmi, C.V. and S.N. Deepa, *Medical Dataset Classification: A Machine Learning Paradigm Integrating Particle Swarm Optimization with Extreme Learning Machine Classifier.* The Scientific World Journal, 2015. **2015**: p. 418060.
5. Yekkala, I., S. Dixit, and M. Jabbar. *Prediction of heart disease using ensemble learning and Particle Swarm Optimization*. in *2017 International Conference On Smart Technologies For Smart Nation (SmartTechCon)*. 2017. IEEE.
6. Annabel, L.S.P., T. Annapoorani, and P. Deepalakshmi. *Machine Learning for Plant Leaf Disease Detection and Classification–A Review*. in *2019 International Conference on Communication and Signal Processing (ICCSP)*. 2019. IEEE.
7. Beyene, T.J., et al., *Assisting differential clinical diagnosis of cattle diseases using smartphone-based technology in low resource settings: a pilot study.* BMC veterinary research, 2017. **13**(1): p. 323.
8. Müller, A.C. and S. Guido, *Introduction to machine learning with Python: a guide for data scientists*. 2016: " O'Reilly Media, Inc.".
9. Burges, C.J., *A tutorial on support vector machines for pattern recognition.* Data mining and knowledge discovery, 1998. **2**(2): p. 121-167.
10. Boswell, D., *Introduction to support vector machines.* Departement of Computer Science and Engineering University of California San Diego, 2002.
11. Géron, A., *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow*. 2022: " O'Reilly Media, Inc.".
12. Raschka, S., *Python machine learning*. 2015: Packt publishing ltd.
13. Mitchell, T.M. and T.M. Mitchell, *Machine learning*. Vol. 1. 1997: McGraw-hill New York.
14. Quinlan, J.R., *Induction of decision trees.* Machine learning, 1986. **1**(1): p. 81-106.
15. Bishop, C.M. and N.M. Nasrabadi, *Pattern recognition and machine learning*. Vol. 4. 2006: Springer.
16. Hastie, T., et al., *The elements of statistical learning: data mining, inference, and prediction*. Vol. 2. 2009: Springer.
17. Géron, A.l., *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow [internet resource] : concepts, tools, and techniques to build intelligent systems*. Second edition.. ed. 2019: Sebastopol, CA : O'Reilly.
18. Dietterich, T.G. *Ensemble methods in machine learning*. in *International workshop on multiple classifier systems*. 2000. Springer.
19. Kuhn, M. and K. Johnson, *Applied predictive modeling*. Vol. 26. 2013: Springer.
20. Li, J., et al., *Feature Selection.* ACM Computing Surveys, 2018. **50**(6): p. 1-45.
21. Saito, T. and M. Rehmsmeier, *The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets.* PloS one, 2015. **10**(3): p. e0118432.

Update the Postgraduate Certificate in Research (PGR) Professional Development Credit



	RD901 Researcher Knowledge and Intellectual Abilities	RD902 Researcher Personal Effectiveness	RD903 Research Governance and Organisation	RD904 Researcher Engagement, Influence and Impact	RD905 Researcher Professional Development Elective	Annual Summary of Credits
2019-2020	Spectacular information and visualising data (1) 202055775 RD901 for EndNote (1) 202055775 RD901 for Endnote - online (1) 202055775 RD901 for Research methods (10) In-sessional English for Academic Purposes courses (1) 202055775 Haddokakon upload evidence	Communicating with confidence: networking skills, assertiveness and understanding personal influence (1) 202055775 RD902 for Interview Techniques for PGRs (1) 202055775 RD902 for PG Essentials (5) 202055775 RD902 for Completing with Confidence - Finishing your PhD on time (1) 202055775 RD902 for upload evidence	Project Management in the Real World (3) 202055775 RD903 for upload evidence	Fear of the Blank Page (1) 202055775 RD904 for Creating Effective Poster Presentations (Science) (1) 202055775 RD904 for Introductory Training for PGRs who Teach (1) file upload required upload evidence	No credit activities	0 (20 pending)
2020-2021	In-sessional English for Academic Purposes courses (4) 202055775 Haddokakon upload evidence	Postgraduate Research Student Induction (1) 202055775 902 2 Post How to be an Effective Researcher (1) 202055775 902 2 How upload evidence	No credit activities	Writing with Confidence - Year 1 (1) 202055775 904 2 Wri Writing with Confidence - Year 2 (1) 202055775 904 2 Wri Communicating with Confidence: Public Speaking for PGRs - Intermediate (1) 202055775 904 2 Comm Writing with Confidence: Getting Going with Academic Writing (1) 202055775 904 2 Wri Communicating with Confidence: Public Speaking for PGRs - Beginners (1) 202055775 904 2 Comm Getting Published in Academic Journals (1) 202055775 904 2 Get upload evidence	No credit activities	0 (12 pending)
2021-2022	No credit activities	Communicating with confidence: Nerves, resilience and pro-resilience for your PhD, your life and your career (1) 202055775 RD902 for upload evidence	Creating Research Data Management Plans (1) 202055775 RD903 2021 Risk Management in the Real World (2) 202055775 RD903 for upload evidence	Writing with Confidence - Year 3 (1) 202055775 RD904 for upload evidence	Writing and Presenting Research (10) 202055775 RD905 Dis upload evidence	0 (15 pending)
2022-2023	No credit activities add activity	No credit activities add activity	Researchers Guide to Ethics (1) PG Essentials: Digital Scholarship Skills (5) file upload required upload evidence add activity	No credit activities add activity	No credit activities add activity	0 (6 pending)
Class Total	Approved: 0 Pending: 18 Total: 18/20	Approved: 0 Pending: 11 Total: 11/10	Approved: 0 Pending: 12 Total: 12/10	Approved: 0 Pending: 10 Total: 10/10	Approved: 0 Pending: 10 Total: 10/10	0 (61 pending)

Grant chat

No	Task		2^{nd}-year student												3^{rd}-year student												4^{th}-year student											
			10/20	11/20	12/20	01/21	02/21	03/21	04/21	05/21	06/21	07/21	08/21	09/21	10/21	11/21	12/21	01/22	02/22	03/22	04/22	05/22	06/22	07/22	08/22	09/22	10/22	11/22	12/22	01/23	02/23	03/23	04/23	05/23	06/23	07/23	08/23	09/23
1	Develop a supervised classification diagnosis methodology of an imbalanced disease data using machine learning (ML) concepts		░	░	░	░	░	░																														
	1.1	Define all key parameters for whole processes		🟩	🟩																																	
	1.2	Develop a proposed methodology			🟩	🟩	🟩	🟩																														
2	Improve a Python-based programme for designing any types of disease diagnosis data									░	░	░	░	░	░																							
	2.1	Define the parameters for proposed model																																				
	2.2	Apply the proposed methodology into data set							🟩	🟩	🟩	🟩	🟩																									
	2.3	Validate the proposed model										🟩	🟩	🟩																								
	2.4	Revise the proposed model												🟩	🟩																							
3	Evaluate the performance of an imbalance disease data methodology for diagnosis using the proposed programme															░	░	░																				
	3.1	Design and revise the model														🟩																						
	3.2	Evaluate the performance															🟩	🟩																				
4	Investigation the expert judgment and combine with the model																															░	░	░	░	░	░	░
	4.1	Design the experiments (if any)																														🟩						
	4.2	Model training																															🟩	🟩	🟩			
	4.2	Verify all parameter and revise																																	🟩	🟩		
	4.3	Analysis result and discussion																																			🟩	
5	Expanding the empirical investigation by applying various diagnosis data and adjusting the model																							░	░	░	░	░									
	5.1	Data preparation and model Training																						🟩	🟩	🟩											
	5.2	Evaluate and compare the results																								🟩	🟩	🟩									
	Milestones and Deliverables																																					
	Annual review report													2																		🟧						
	Mini VIVA														1																							
	Paperwork for each year and Publication												1												2								🟧			🟧		
	3^{rd} Year progress meeting																										1					🟧						
	All work to be completed																																					🟧
	Thesis submission																																					
	Course end date																																					