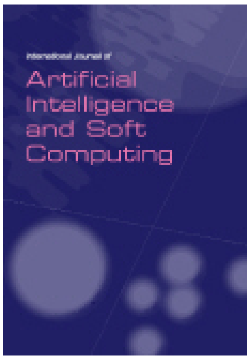



**International Journal of Artificial Intelligence and Soft Computing**

ISSN online: 1755-4969 - ISSN print: 1755-4950
https://www.inderscience.com/ijaisc

# PigB: intelligent pig breeds classification using supervised machine learning algorithms

Pritam Ghosh, Satyendra Nath Mandal

**DOI:** 10.1504/IJAISC.2022.10051226

**Article History:**

| | |
|---|---|
| Received: | 02 July 2022 |
| Last revised: | 09 August 2022 |
| Accepted: | 09 August 2022 |
| Published online: | 21 October 2022 |

# PigB: intelligent pig breeds classification using supervised machine learning algorithms

Pritam Ghosh*

Department of Computer Science and Technology,
Iswar Chandra Vidyasagar Polytechnic,
Sevayatan, Jhargram, West Bengal, 721514, India
Email: ghoshpritam25@gmail.com
*Corresponding author

Satyendra Nath Mandal

Department of Information Technology,
Kalyani Government Engineering College,
Kalyani, Nadia, West Bengal, 741235, India
Email: satyen_kgec@rediffmail.com

**Abstract:** Different supervised machine learning algorithms' performance varies when applied to different datasets. Moreover, using a generalised supervised algorithm may not be able to produce the optimal performance as the nature of data is different for different datasets. In this paper, we used a pig breed dataset containing various statistical features extracted from individual pig images of five pig breeds. Eight well-established algorithms such as logistic regression, multilayer perceptron, decision trees, gradient boosted decision trees, random forest, support vector machine, K-nearest neighbours, and naïve Bayes are carefully applied to this dataset by tuning the necessary hyperparameters for each algorithm. The performance of all the applied algorithms is compared based on the micro averaged area under precision-recall (PR) and receiver operator characteristics (ROC) curves. From the results obtained, the SVM algorithm with a radial basis function (RBF) kernel has outperformed all the other algorithms with a pig breed prediction accuracy of 98%.

**Keywords:** supervised learning; fine grained classification; statistical features; performance evaluation; hyperparameter optimisation; receiver operator characteristics; ROC; support vector machine; SVM; radial basis function; RBF.

**Reference** to this paper should be made as follows: Ghosh, P. and Mandal, S.N. (2022) 'PigB: intelligent pig breeds classification using supervised machine learning algorithms', *Int. J. Artificial Intelligence and Soft Computing*, Vol. 7, No. 3, pp.242–266.

**Biographical notes:** Pritam Ghosh has earned his MTech from the Kalyani Government Engineering College in 2020. He is currently working as a Lecturer in the Department of Computer Science and Technology in Iswar Chandra Vidyasagar Polytechnic, West Bengal. He has published five research papers in peer-reviewed international journals and six papers published in reputed national and international conferences. He received the best research paper award in International Conference on Frontiers in Computing and Systems. His research interests include IoT, blockchain, and machine learning.

Satyendra Nath Mandal completed his PhD from the Maulana Abul Kalam Azad University of Technology, Kolkata, India and is currently an Assistant Professor in the Department of Information Technology of Kalyani Government Engineering College. He has authored more than 100 research papers including conferences and journals and is currently a principal investigator in several government funded research projects. His research interests include soft computing, fuzzy logic, artificial intelligence, and intelligent biometrics.

## 1 Introduction

One of the most promising but difficult issues in intelligent livestock management is the accurate and reliable identification of domestic animal breeds from photographs. Classification of an animal's breed is essential for preserving the integrity of the breed to which the animal belongs. Maintaining breed purity for economically valuable domestic animals, such as pigs, is crucial since the value of the animal is directly proportional to its breed. In the north-eastern region of India, pig farming is among the main industries. According to the 20th Livestock Census, India has 9.06 million pigs, which accounts for 2.01% of its total livestock (Ministry of Fisheries, Animal Husbandry & Dairying, 2019). Animal husbandry and the livestock industry play a significant part in the socioeconomic development of a nation. According to the 2008 World Development Report, the livestock industry is responsible for 70% of rural economic development (Pica et al., 2008). Breed identification is also required for implementing the Global Plan of Action for Animal Genetic Resources (AnGR) as outlined by the Food and Agriculture Organization of the United Nations (FAO) (Hoffmann and Scherf, 2010).

Phenotypic and genotypic characterisation is regarded as the gold standard for breed identification. Phenotypic characterisation includes identification, quantitative and qualitative description, documenting of populations, knowledge of natural environments, and production processes (Food and Agriculture Organization of the United Nations, 2012) of a breed. Genetic characterisation is performed to comprehend the variety and uniqueness of a genetic resource for the goal of formulating policies to enhance the resource's worth and breadth of use (Food and Agriculture Organization of the United Nations, 2011). However, phenotypic characterisation is reliant on animals' habitat and diet, and defining phenotypic characteristics is time-consuming. However, genotypic characterisation is highly expensive. To characterise the genotype of animals, specialised labs are required. Therefore, these current breed identification systems are unsuitable for application in scenarios requiring immediate and accurate breed prediction in the field. In real farming situations, livestock owners rely on their own observational abilities to identify breeds. This method, however, has a major drawback because, according to various veterinary researchers, the interbreed mixing is so high in the pig populations of India that it is extremely difficult to identify the pure breed genetic resources, even with years of experience in livestock maintenance, thus increasing the likelihood of misclassification. This has a negative impact on breed conservation, which is highly alarming given that India is one of the few nations that contribute significantly to the worldwide cattle gene pool. Therefore, a new approach for animal breed identification is required that is accurate, economically feasible, and accessible to the general public. This was the impetus for this investigation.

A collection of physical characteristics may be used to identify breeds of domestic animal species. These characteristics may be retrieved from an animal's photograph. The purpose of this research is to build a method for automatically identifying pig breeds from photographs using supervised machine learning trained on a collection of morphological traits expressed numerically as statistical parameters and colour components. This automated approach is aimed to solve the difficulties associated with genotypic, phenotypic, and manual identification of pig breeds.

Using the camera on a smart phone, photographs of five distinct pig breeds were taken on farms that raise only pure-bred pigs. To aid with error-free segmentation and feature extraction, all individual pig photos have been captured against a consistent green backdrop. In order to generate the pig breed dataset, the collected pictures were segmented and ten statistical and colour component characteristics were extracted from the segmented images. In diverse datasets, several well-known supervised algorithms have previously delivered accurate classifications. Given the introduction of the pig breed statistics and colour components dataset in this research, it is unknown which supervised algorithm will perform best on this data, given that supervised classification methods are data-driven. Extensive research in the literature has shown that no one supervised classification method can produce the best results for all applications (Table 1). Finding the supervised classification algorithm that may generate the best results for this dataset is thus a crucial challenge. Therefore, instead of using a general classification algorithm, we applied eight distinct well-established algorithms with suitable hyperparameter tweaking to the pig breed dataset and picked the approach with the greatest overall performance based on this evaluation. The performance of each algorithm was examined using the AUC ROC and AUC PR metrics.

## *1.1 Contributions*

- It has been observed that SVM with RBF kernel is the most capable supervised classification algorithm for image-based pig breed prediction with 98% accuracy.
- A new pig breed dataset has been introduced containing a total of 250 individual pig images from five different breeds.
- Established the possibility that pig breeds can be automatically classified based on their statistical properties and colour components extracted from the contents of individual pig images.

The following paper is broadly divided into seven sections. Starting with introduction in Section 1, the related works is discussed in Section 2, followed by data acquisition and dataset creation in Section 3. The pig breed image dataset details are furnished in Section 3 and the process of extracting statistical data from the dataset is given in Section 4. The methodology is discussed in Section 5, having details of all the algorithms used in this paper followed by implementation details and hyperparameter optimisation in Section 6. The performance evaluation techniques, prediction results and comparison of all the algorithms are discussed in Section 7, followed by conclusion in Section 8.

**Table 1** Performance of different supervised algorithms in different applications (see online version for colours)

| *Area of application* | *Logistic regression* | *Multilayer perceptron* | *Decision tree* | *Random forest* |
|---|---|---|---|---|
| Mosquito identification from backscattered optical signals (Genoud et al., 2020) | X | X | 85.70% | X |
| Extracting useful information from unstructured web data (Sheshasaayee and Thailambal, 2017) | X | X | X | 88.00% |
| Diagnosing disease states by classifying immunosignaturing data (Kukreja et al., 2012) | 78.90% | 87.30% | X | 70.00% |
| Recognising specific arm positions for telerobotic control using electromyography signals (Frasca et al., 2016) | X | X | 95.10% | 97.30% |
| Classification of bat echolocation calls (Armitage and Ober, 2010) | X | 67.00% | X | 85.00% |
| Diabetes prediction (Osisanwo et al., 2017) | X | X | 63.20% | 65.30% |
| *Area of application* | *Support vector machine* | *Naïve Bayes* | *Gradient boosted decision trees* | *K-nearest neighbours* |
| Mosquito identification from backscattered optical signals (Genoud et al., 2020) | 88.80% | 83.30% | X | 81.50% |
| Extracting useful information from unstructured web data (Sheshasaayee and Thailambal, 2017) | 97.40% | 90.20% | X | X |
| Diagnosing disease states by classifying immunosignaturing data (Kukreja et al., 2012) | 87.00% | 90.40% | X | 76.20% |
| Recognising specific arm positions for telerobotic control using electromyography signals (Frasca et al., 2016) | 95.70% | 90.00% | X | 97.80% |
| Classification of bat echolocation calls (Armitage and Ober, 2010) | 70.00% | X | X | X |
| Diabetes prediction (Osisanwo et al., 2017) | 74.00% | 67.80% | X | X |

## 2 Background study

Different phenotypic characteristics, such as muzzle print, body form, coat colours and pattern, have been used to identify animal breeds (Kumar et al., 2018; Andrew et al., 2016; Lahiri et al., 2011). Genotype-based marking techniques, such as whole-genome sequencing, microsatellite markers, and DNA barcoding, have also been used for animal breed identification (Wang et al., 2016; Sardina et al., 2015; Peng et al., 2019). Diverse research (Hailu, 2015; Kumar et al., 2018; Burghardt, 2012) have emphasised the difficulties in phenotypic and genotypic breed identification and the possibilities of using morphological characteristics in animal breed identification.

Computer-based animal breed identification is an active area of study, and several approaches have been developed for various species of animals, including cats, dogs, goats, cattle, kangaroos, and sheep, among others (Yu et al., 2013; Liu et al., 2016; Zhang et al., 2015; Sundaram and Loganathan, 2020). CNNs have recently become the most effective technique for any image-based classification assignment. Consequently, many CNN types have been used for animal breed categorisation (Parkhi et al., 2012; Meena and Agilandeeswari, 2019; Ayanzadeh and Vahidnia, 2018; Jwade et al., 2019; Mandal et al., 2020). The main disadvantage of employing CNN-based approaches is the enormous amount of time required to effectively train these models (He and Sun, 2015).

Various learning-based intelligent optimisation algorithms have been studied, such as elephant herding optimisation algorithm and DLEA (Li et al., 2021a, 2021b, 2020b). These algorithms combine learning different learning mechanisms to improve the learning ability and provide better optimisation. On a variety of challenging classification problems, supervised machine learning methods such as decision tree (DT), SVM, logical regression, multilayer perceptron (MLP), etc. have been used. In several investigations, these classification techniques have been used (Table 1).

In Table 1, 'X' entry denotes that the algorithms have not been applied on the problem. The coloured cell represents the best performing algorithms along with their accuracies. From this we can clearly observe that a single algorithm cannot perform optimally for all the different tasks. These algorithms take very less time to train but are not capable of directly working on images. It is also uncertain that the results may have changed if the unused algorithms were also used for those cases where it has not been used.

## 3 Dataset

### 3.1 Image capturing

This article predicts the pig breeds based on their statistical and colour characteristics. These characteristics were taken from photos of individual pigs. Among the various pig breeds, five have been selected for study: Ghungroo, Mali, Hampshire, Duroc, and Yorkshire (Figure 1).

**Figure 1** Sample images of individual animals from five pig breeds with uniform background (see online version for colours)

The images of individual animals from the breeds were collected from organised farms maintained by India's leading research institutes:

1 ICAR National Research Centre on pig, Rani, Assam.

2 ICAR Research Complex for NEH Region, Umiam, Meghalaya.

3 ICAR Research Complex for NEH Region, Tripura Centre, Tripura.

During the process of catching individual pigs, several limits have been introduced to decrease the effect of the environment. The following measures have been taken to capture the pigs:

1 A green drape was hung over the pigs' standing area and three sides to provide a homogeneous backdrop.

2 The cell phone was positioned on one side parallel to the pigs and about 2 metres away.

3 The moveable lens was positioned about in the middle of the animal's length and perpendicular to the median sagittal plane.

4 The lens was positioned a few inches below the animal's height in order to capture all visible sections of a pig.

5 The photographs were taken using natural light and a side profile angle.

6 Figure 2 depicts how photos of pigs were collected such that the right or left side of their bodies are plainly apparent.

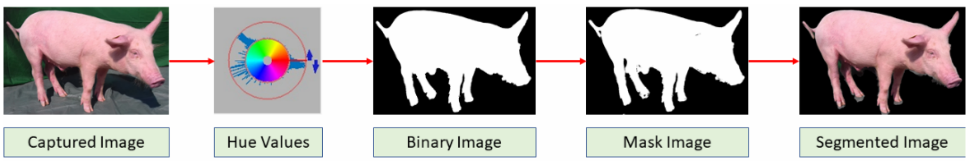
**Figure 2** Capturing the pig images (see online version for colours)

## 3.2 *Pig image segmentation*

We made the collected photos uniform in size. Every pixel was transformed to the HSV colour model. The hue was obtained by holding the saturation and value constant while omitting the background from each photograph of a pig breed used for the research. The

HSV picture was then transformed to a binary image using the hue value. The visible portions of the pig in the original image became white in the binary image, while the backdrop became black. The resulting black and white binary picture may include some white blobs. The regions of the white spots were determined. To create a mask picture, all blobs were inverted to black except for the biggest blob. As seen in Figure 3, the generated mask picture was superimposed on the original RGB image to produce the segmented image of the pig.

**Figure 3** Hue calculation and segmentation of pig images (see online version for colours)




## 4 Statistical parameters and colour components extraction

From the segmented photos, numerous statistical measures on the content and colour component of images are generated to assist define the characteristics of certain breeds. For the purpose of obtaining statistical parameters, the segmented pictures were converted to greyscale and stored in memory as a 2D matrix. The hue, saturation, and value components of the HSV colour model have been retrieved from the RGB picture by converting it to an HSV image (Chernov et al., 2015). HSV colour characteristics are crucial for sensing visual surroundings, recognising objects, and communicating information (Arivazhagan et al., 2013). HSV colour model is resistant to external lighting changes. Using HSV colour model ensures the colour of the captured images remain consistent even if there are minor changes in the lighting conditions while capturing the image. Subsequently data extraction from the images will be consistent resulting in better classification.

Texture is one of the most significant properties for classifying and identifying things (Haralick et al., 1973). The numerous statistical metrics generated from an image may be used to define an image’s texture. In this research, seven of these characteristics, including entropy, standard deviation, mean, sum, max, min, variance, median, and mode, have been determined using the MATLAB environment (Gonzalez et al., 2003). MATLAB version 2015a installed on a system with 2.3 GHz Intel Xeon processor, 64 GB RAM, 2 TB 7200 rpm hard disk and Nvidia Quadro P4000 graphics card with 8 GB video memory is used for performing all the necessary calculations. After all parameters have been calculated, they are saved in a comma-separated values file. The CSV file includes information for 250 individual pigs, including 50 pigs from each of the five breeds listed in Table 2.

**Table 2** Statistical parameter and colour component values (sample data for two pigs from each breed) (see online version for colours)

| *Sl. no.* | *Image ID* | *H* | *S* | *V* | *Entropy* | *Std. dev.* | *Mean* | *Sum* | *Median* | *Mode* | *Variance* | *Breed* | *Class value* |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | G1 | 0.152 | 0.107 | 0.074 | 2.374 | 24.748 | 16.342 | 1,863.030 | 0.000 | 1.308 | 6,519.171 | Ghungroo | 3 |
| 2 | G2 | 0.145 | 0.105 | 0.074 | 2.290 | 24.628 | 16.313 | 1,859.722 | 0.000 | 1.278 | 6,635.538 | Ghungroo | 3 |
| 3 | Y1 | 0.336 | 0.253 | 0.255 | 4.132 | 56.091 | 53.229 | 6,068.157 | 0.096 | 31.222 | 9,582.595 | Yorkshire | 1 |
| 4 | Y2 | 0.278 | 0.236 | 0.203 | 3.516 | 55.982 | 41.956 | 4,782.939 | 0.025 | 16.606 | 8,683.902 | Yorkshire | 1 |
| 5 | M1 | 0.142 | 0.062 | 0.141 | 3.599 | 45.877 | 35.085 | 3,999.646 | 0.000 | 10.606 | 4,572.518 | Mali | 5 |
| 6 | M2 | 0.104 | 0.047 | 0.112 | 3.035 | 35.341 | 28.077 | 3,200.798 | 0.000 | 5.409 | 4,694.001 | Mali | 5 |
| 7 | H1 | 0.224 | 0.230 | 0.159 | 4.112 | 36.803 | 33.704 | 3,842.253 | 0.040 | 19.449 | 10,009.343 | Hampshire | 4 |
| 8 | H2 | 0.191 | 0.196 | 0.153 | 4.120 | 36.020 | 32.382 | 3,691.535 | 0.010 | 18.854 | 9,453.684 | Hampshire | 4 |
| 9 | D1 | 0.073 | 0.259 | 0.259 | 4.092 | 56.073 | 52.393 | 5,972.768 | 0.010 | 33.924 | 9,889.457 | Duroc | 2 |
| : | : | : | : | : | : | : | : | : | : | : | : | : | : |
| 250 | D2 | 0.078 | 0.305 | 0.265 | 4.226 | 58.881 | 52.341 | 5,966.879 | 0.040 | 27.641 | 8,910.063 | Duroc | 2 |

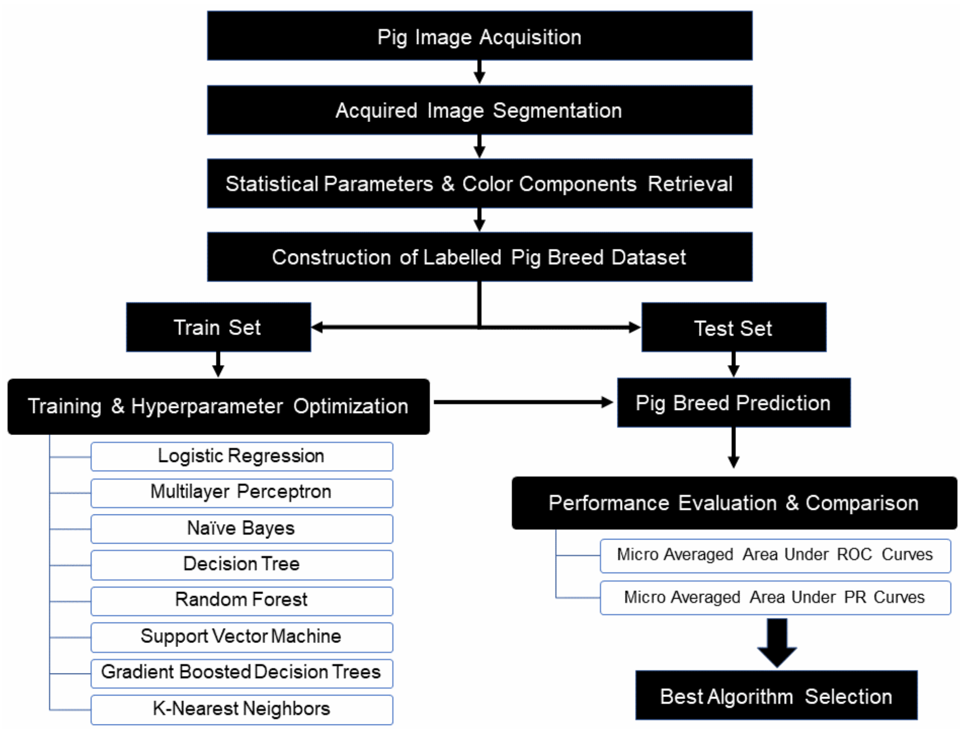
## 5 Methodology

In this study, eight supervised classification methods were used to a statistical and colour component-based pig breed dataset in order to forecast pig breeds. The nest algorithm has been chosen based on its predictive accuracy. The ideal algorithm selection procedure has been broken into a series of phases that function as a layered structure (Figure 4). The algorithms utilised in this investigation are detailed in the next section.

### *5.1 Logistic regression (LR)*

The classification procedure of LR is used to assign data to a discrete set of classes. In contrast to linear regression, LR changes its output using the logistic sigmoid function to yield a probability value that may then be translated to two or more discrete classes (Kleinbaum et al., 2002).

**Figure 4** Plan of action (see online version for colours)




To train a good LR classifier we must prevent it from overfitting. This is controlled by the regularisation parameter which controls how closely the model fits to the training data. Also, the type of training algorithm used (optimiser) plays an important role for getting a good LR classifier. We have tested three different optimisers such as L-BFGS (Liu and Nocedal, 1989), LIBLINEAR (Fan et al., 2008), Newton-CG (Bollapragada et al., 2019) and ten different regularisation parameter values (Table 3).

## *5.2 Multilayer perceptron*

MLP is an artificial feed-forward neural network. This method is composed of many linked processing pieces known as perceptron (artificial neurons) (Ruck et al., 1990). Several hyperparameters affect the training of an effective MLP classifier. In this study, we evaluated the most significant variables, including hidden layer sizes, activation functions [logistic, Tanh, ReLU (Karlik and Olgac, 2011)], and optimisers [L-BFGS (Liu and Nocedal, 1989), SGD (Bottou, 2010), and Adam (Kingma and Ba, 2014)]. To prevent overfitting, several L2 regularisation penalty values, the number of training iterations, and different strategies for updating the learning rate have been evaluated (Table 3).

## *5.3 DT, random forest (RF) and gradient boosted decision trees (GBDT)*

A DT is a supervised learning approach whose structure resembles a flowchart. It has one root node, many internal nodes, and leaf nodes. The root node holds the best characteristic picked among a variety of attributes depending on certain criteria. Except for leaf nodes, each node has two outgoing paths or branches. In each internal node, the test is executed, and each branch is represented by the result of one test. Finally, all pathways terminate in leaf nodes, which are nothing more than the problem's class label (Loh, 2011). The most crucial hyperparameter for a DT is its depth, since depth is closely connected to DT classifier overfitting. We trained the DT classifier with six distinct depth settings (Table 3).

The RF classifier consists of many individual DTs that operate as an ensemble. Each individual tree in the RF predicts the class based on input values and final output is given by majority voting (Liaw and Wiener, 2002). Like a DT, RF classifier overfitting can be prevented my controlling the max depth of each tree in the forest and the total number of such trees in the forest (Table 3).

GBDT is an ensemble method that aggregates several simple decision trees to create on strong decision tree by effectively learning from its mistakes with each iteration (Friedman, 2002). To reduce overfitting in this algorithm, we have tested three different hyperparameters, such as learning rate, number of boosting stages (total number of trees) and the depth of each individual tree (Table 3).

## *5.4 Support vector machine*

The purpose of the SVM method is to classify data points using a hyperplane in an n-dimensional space (where n is the number of characteristics). There are several potential hyperplanes that might be selected to divide the two classes of data points. Our goal is to locate the plane with the greatest margin, or the greatest distance between data points of both classes. Maximising the margin distance gives reinforcement so that subsequent data points may be identified with more certainty (Hearst et al., 1998). SVMs translate data that is not linearly separable in n-dimensional space into a higher dimension where it is linearly separable using the kernel approach. Therefore, choosing the proper kernel function is essential. We have evaluated three distinct kernels, including linear, polynomial, and RBF (Pilario et al., 2020) as well as various regularisation values for decreasing overfitting (Table 3).

### 5.5 Naïve Bayes (NB)

NB is a supervised learning algorithm based on applying Bayes' theorem with the 'naïve' assumption of conditional independence between every pair of features given the value of the class variable (Rish, 2001). In this paper, we have used the Gaussian NB algorithm, where the likelihood of the features is assumed to be Gaussian (Chan et al., 1982).

### 5.6 K-nearest neighbours (KNN)

KNN is one of the simplest algorithms used in machine learning for multiclass classification. KNN algorithms use data and classify new data points based on similarity measures (e.g., distance function). Classification is done by a majority voting to its neighbours. The data is assigned to the class which has the nearest neighbours (Chomboon et al., 2015). In this paper, we have tested five different K values which represent the number of nearest neighbours to include in most of the voting process (Table 3).

## 6 Implementation

We have experimented with eight different supervised classification algorithms with the objective of finding the most accurate algorithm for pig breed classification using the data from the newly acquired dataset. Classification has been performed between five different pig breeds making it a multiclass classification task. To gain knowledge about the algorithm's performance on each of the breeds, one-vs.-rest strategy has been used (Figure 5). In this method, if there are total five classes then we need to train five different classifiers, where each class is fitted against all the other classes. This method is very easy to interpret and evaluate because it converts the multiclass classification problem to $n$ different binary classification problems (where $n$ is the total number of classes). In our case, there are five pig breeds, and eight different supervised algorithms, therefore 40 classifiers have been trained and evaluated.

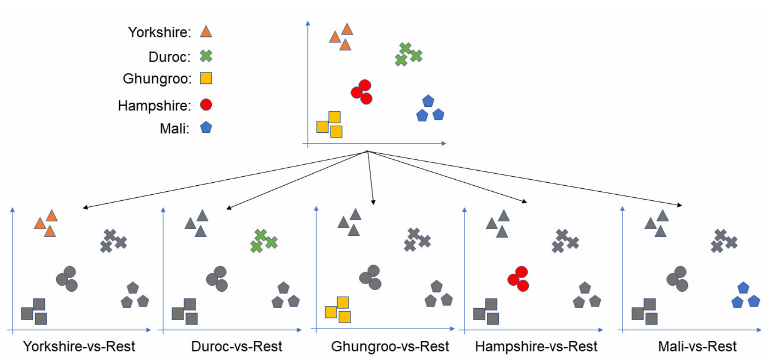
Different algorithms have different hyperparameters and have different effect on the performance of the classifier. Hyperparameters are the important parameters which need to be tuned manually so that the algorithm is trained properly for the given task. This is a very important step for building any supervised machine learning classifier. Details of the hyperparameters for each algorithm have been demonstrated in Table 3. The best hyperparameter combination has been selected based on the overall training accuracy for each classifier.

To train each classifier, the total data has been divided into two different sets (training and testing) in 80:20 ratio based on the Pareto principle. After studying different related research works, we found that the 80:20 split ratio is the most common and effective ratio for splitting the dataset into train and test sets. That means, out of the 250 samples (50 samples from five breeds), 200 samples (40 samples from five breeds) have been used for training and 50 samples (ten samples from five breeds) have been used for testing. In this paper we have stuck with the 80:20 split ratio for training and testing all the algorithms. Since the dataset is relatively small, to reduce overfitting five-fold cross-validation has been used during training (Raschka¸2018).

**Table 3** Hyperparameter selection (training accuracy is obtained from the selected set of hyperparameters)

| *Algorithm* | *Hyperparameter* | *Options* | *Selected* | *Maximum training accuracy* |
|---|---|---|---|---|
| Multilayer perceptron | Hidden layer size | 10, 50, 100 | 50 | 57.77% |
| | Activation function | Logistic, Tanh, ReLU | Logistic | |
| | Optimiser | L-BFGS, SGD, Adam | L-BFGS | |
| | L2 penalty | 0.0001, 0.001, 0.01, 0.1, 1, 10, 100 | 0.001 | |
| | Learning rate update | Constant, inverse scaling, adaptive | Constant | |
| | Iterations | 100, 500, 1,000 | 500 | |
| Decision tree | Max depth | 2, 4, 8, 16, 32, None | 16 | 97.00% |
| Random forest | Number of trees in forest | 5, 50, 100, 250 | 50 | 96.33% |
| | Max depth | 2, 4, 16, 32, None | None | |
| Support vector machine | Regularisation parameter | 0.1, 0.5, 0.8, 1, 1.3, 1.5, 2, 2.5, 3 | 0.5 | 98.50% |
| | Kernel | Linear, polynomial, RBF | RBF | |
| Logistic regression | Regularisation parameter | 0.1, 0.5, 0.8, 1, 1.3, 1.5, 2, 2.5, 3, 15 | 0.8 | 67.50% |
| | Optimiser | L-BFGS, LIBLINEAR, Newton-CG | LIBLINEAR | |
| Naïve Bayes | - | - | - | 63.77% |
| K-nearest neighbours | No. of neighbours (K) | 2, 3, 4, 5,6 | 5 | 73.50% |
| Gradient boosted decision trees | Learning rate | 0.01, 0.1, 1, 10, 100 | 0.01 | 88.66% |
| | Boosting stages | 5, 50, 250, 500, 1,000 | 500 | |
| | Max depth | 1, 3, 5, 7, 9 | 3 | |

**Figure 5** One-vs.-rest strategy (for each instance, only one breed is taken as positive class and all others breeds combined is taken as negative class) (see online version for colours)



The details of all the algorithms and their specific hyperparameters have been discussed in the previous section. The complete process has been implemented using Python and Scikit-Learn (Pedregosa et al., 2011) on a system with 2.3 GHz Intel Xeon processor, 64 GB RAM 2 TB 7200 rpm hard disk and Nvidia Quadro P4000 graphics card with 8 GB video memory.

## 7 Results and discussion

All the eught algorithms have been evaluated on the 50 samples in the test set. Two separate performance evaluation metrics have been used to measure the classification performance of those algorithms.

### *7.1 Performance evaluation*

For any classification scenario, the classifier predictions can be divided into four different groups, such as true positive predictions, false positive predictions, false negative predictions and true negative predictions.

- True positive (TP): classifier correctly predicts the positive class.
- False positive (FP): classifier incorrectly predicts the positive class.
- False negative (FN): classifier incorrectly predicts the negative class.
- True negative (TN): classifier correctly predicts the negative class.

From these predictions various metrics such as precision, recall (true positive rate) and false positive rate can be calculated as shown in equation (1), equation (2) and equation (3).

$$\text{Precision} = \frac{TP}{TP + FP} \tag{1}$$

$$\text{Recall(True Positive Rate)} = \frac{TP}{TP + FN} \tag{2}$$

$$\text{False Positive Rate} = \frac{FP}{FP + TN} \tag{3}$$

Furthermore, for predicting the class of any test sample, the classifier produces probability values for both the positive and negative class which sums to unity. Depending on the classification threshold selected, if at least one of the probability values is more than the selected threshold, the test sample is classified to be of that class for which it has occurred. That means, if the classification threshold is 0.5, and the classifier output is (positive class: 0.7, negative class: 0.3), then the prediction belongs to the positive class since its probability is more than the threshold.

If different thresholds are selected, the classifier predictions will vary accordingly. To take this variation into account, two metrics called PR curve (Davis and Goadrich, 2006) and receiver operating characteristic curve (Fawcett, 2006) have been developed.

The dependence of precision, recall (TPR) and FPR on classification threshold is given in Table 4.

**Table 4** Relationship between precision, recall (TPR) and FPR

| *Confidence* | *Precision* | *Recall (TPR)* | *FPR* |
|---|---|---|---|
| High | High (less false positives) | Low (more false negatives) | Low (less false positives) |
| Low | Low (more false positives) | High (less false negatives) | High (more false positives) |

The ROC curve is drawn between FPR in the x-axis and TPR in the y-axis. The PR curve is drawn between recall in the x-axis and precision in the y-axis. Each point on the ROC curve and PR curve represents a specific classification threshold. That means the area under such a curve will be classification-threshold-invariant. Therefore, these two metrics can determine the classifier performance in a neutral scenario.

The AUC PR metric is necessary in our experiments because, we are converting the multiclass problem to a binary problem. What that implies is that our data is becoming imbalanced, where we have fewer positive samples and more negative sample. This is taken care of by the AUC PR metric by comparing the precision and true positive rate rather than comparing the false positive rate with the true positive rate, because a large change in the number of false positives can lead to a small change in the false positive rate.

## *7.2 Performance comparison*

For all the algorithms, for each pig breed both the metrics (AUC PR and AUC ROC) have been calculated. To get the overall performance for all the breeds, the breed wise performance has been combined using micro-averaging. Micro-average aggregates the contributions of all the classes to compute the average metric. For both the metrics, if the area under the curve (AUC) for a classifier is close to unity, then the classifier is said to be more accurate. That means more AUC means better classifier. The PR plots and ROC plots for each algorithm are shown in Figures 6–13. The micro-averaged AUC PR and AUC ROC is tabulated in Table 5.

**Table 5** Performance comparison (see online version for colours)

| *Algorithm* | *Micro avg. AUC PR* | *Micro avg. AUC ROC* |
|---|---|---|
| Decision tree | 0.4604 | 0.7625 |
| K-nearest neighbours | 0.5284 | 0.8173 |
| Support vector machine | 0.8600 | 0.9782 |
| Random forest | 0.8160 | 0.9125 |
| Naïve Bayes | 0.3179 | 0.7021 |
| Multilayer perceptron | 0.4432 | 0.7616 |
| Logistic regression | 0.7233 | 0.9170 |
| Gradient boosted decision trees | 0.5960 | 0.8300 |

Plotting out the data in Table 5, from highest performing algorithm to lowest performing algorithm, for both the metrics (Figure 14) reveals that SVM algorithm outperforms all the others by a significant margin.

**Figure 6** PR and ROC for RF (see online version for colours)

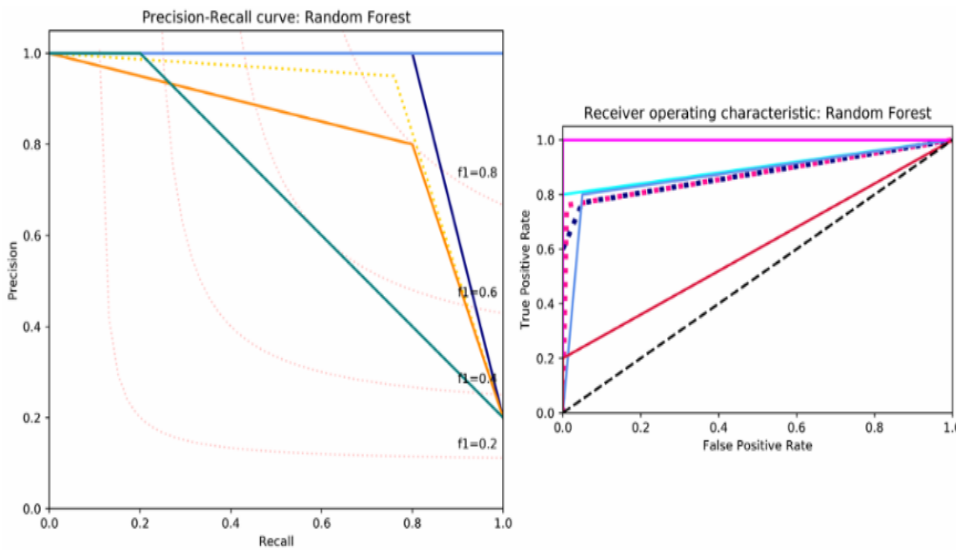



**Figure 7** PR and ROC for NB (see online version for colours)

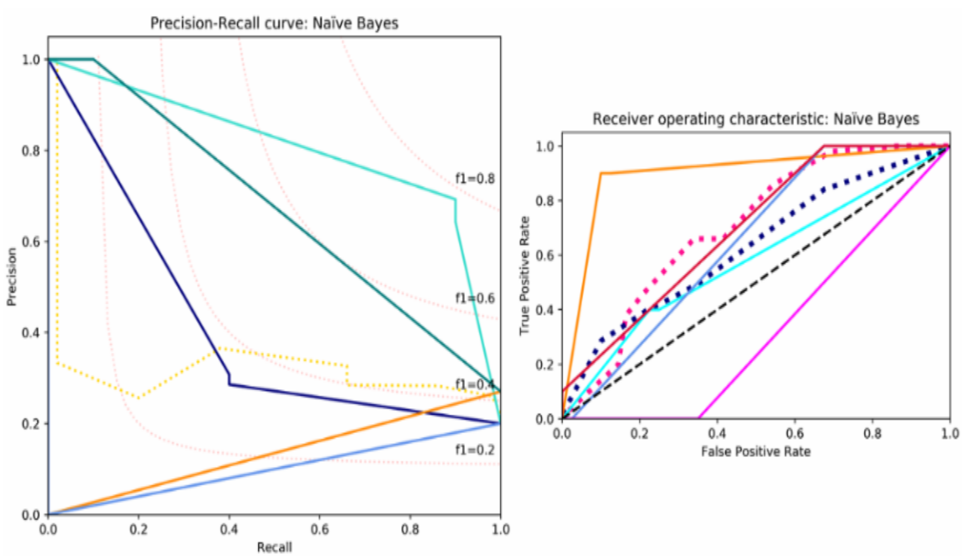


**Figure 8** PR and ROC for MLP (see online version for colours)

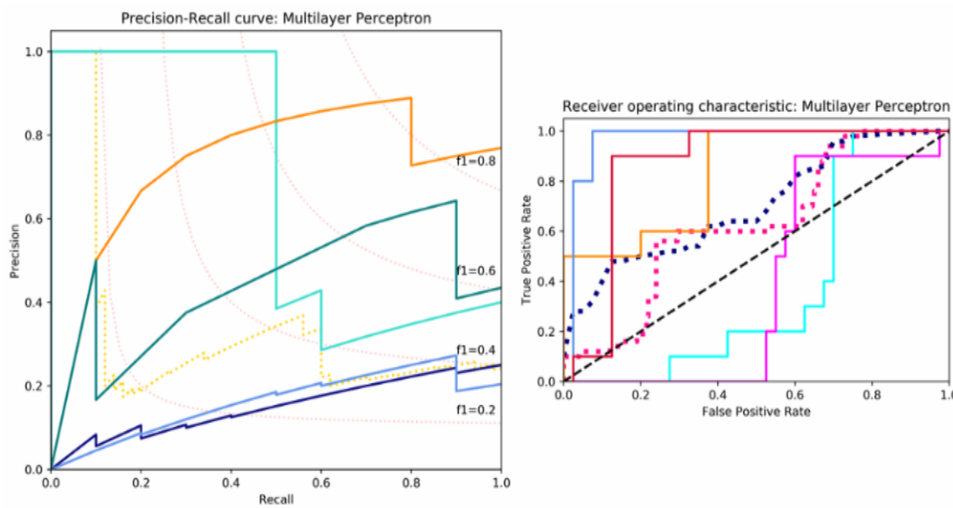
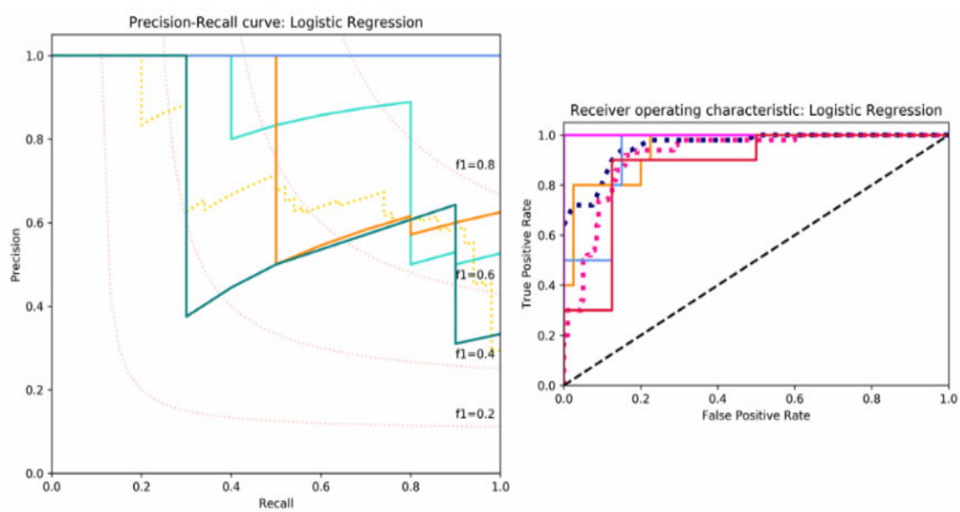



**Figure 9** PR and ROC for LR (see online version for colours)




## 7.3 *Performance comparison of different supervised algorithms in different classification scenarios*

The trained SVM classifier is used to also predict the breeds for the 50 samples in the test dataset for calculation the classifier accuracy. The prediction results and the overall accuracy are shown as a confusion matrix in Figure 15. This pig breed prediction accuracy has been compared with other research done with supervised algorithms as shown in Table 6. From this comparison classifier developed in this paper produces better results compared to the rest.

Based on the results obtained from Figure 14, we can conclude that SVM with RBF kernel outperforms the rest of the algorithms in case of pig breed classification. We then used this trained SVM classifier on a set of pig breed images containing ten images from each breed. The confusion matric for the classification output is shown in Figure 15.

**Figure 10** PR and ROC for KNN (see online version for colours)

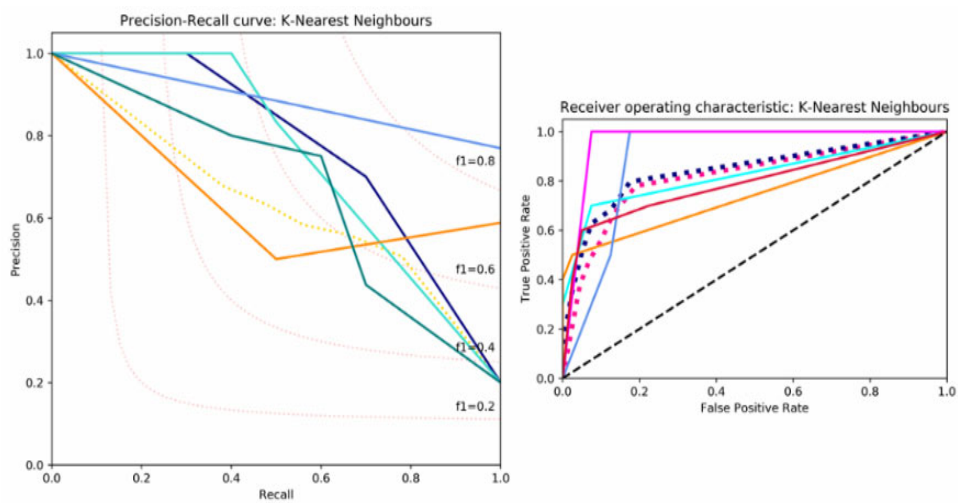



**Figure 11** PR and ROC for gradient boosted decision (see online version for colours)

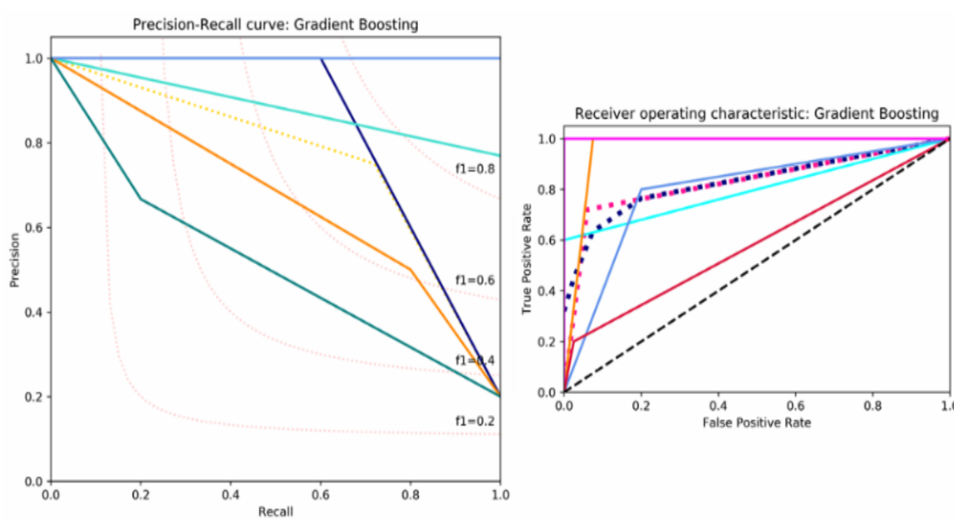


**Figure 12** PR and ROC for DT (see online version for colours)

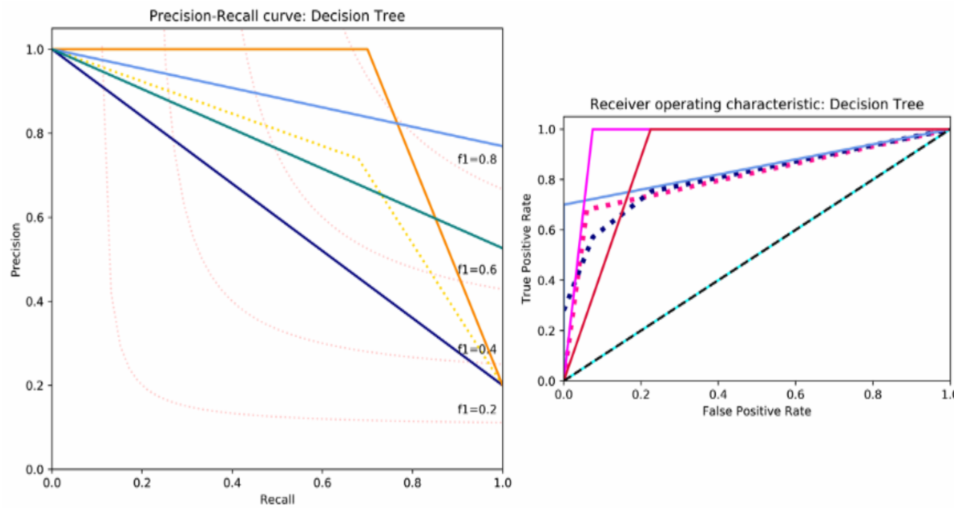
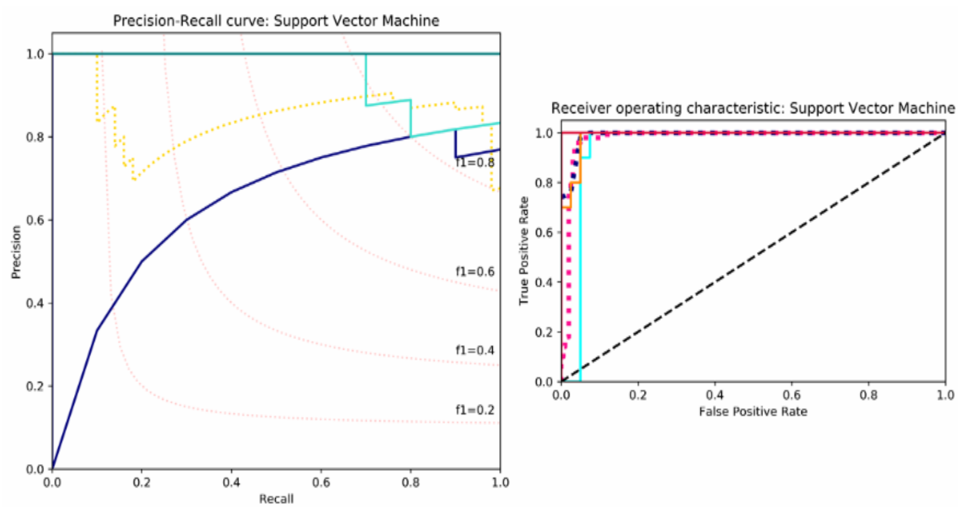



**Figure 13** PR and ROC for SVM (see online version for colours)




From Figure 15, we can clearly observe that only in case of the Yorkshire breed the trained SVM classifier misclassified one sample image as Duroc breed. But in all other cases it produced excellent accurate results without any misclassification producing an overall breed accuracy of 98%. This result is compared to various other multiclass classification problems in Table 6.

**Figure 14** Plot of performance of the algorithms in decreasing order (see online version for colours)

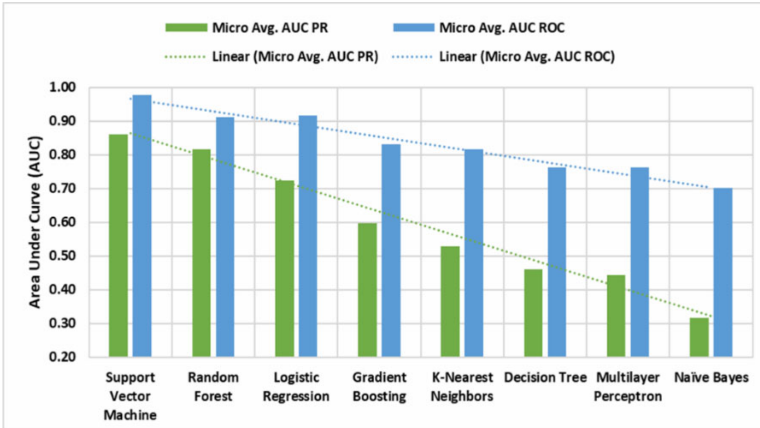


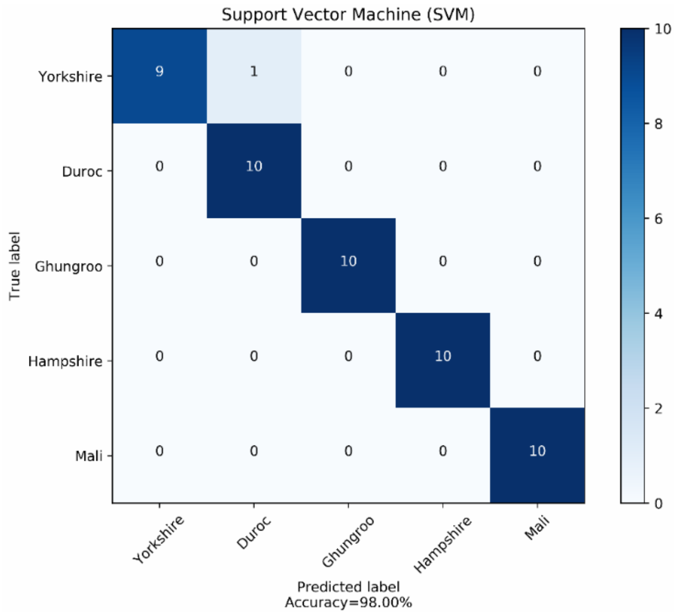

**Table 6** Accuracy comparison (see online version for colours)

| *Area of application* | *Best algorithm* | *Accuracy* |
|---|---|---|
| Mosquito identification from backscattered optical signals (Genoud et al., 2020) | Support vector machine | 88.80% |
| Extracting useful information from unstructured web data (Sheshasaayee and Thailambal, 2017) | Support vector machine | 97.40% |
| Diagnosing disease states by classifying immunosignaturing data (Kukreja et al., 2012) | Naïve Bayes | 90.40% |
| Recognising specific arm positions for telerobotic control using electromyography signals (Frasca et al., 2016) | K-nearest neighbours | 97.80% |
| Classification of bat echolocation calls (Armitage and Ober, 2016) | Random forest | 85.00% |
| Diabetes prediction (Osisanwo et al., 2017) | Support vector machine | 74.00% |
| Pig breed prediction (This paper) | Support vector machine | 98.00% |

## 8 Conclusions

In this paper, we have demonstrated that pig breed classification can be done from statistical parameter and colour component features extracted from images. Individual pig images from five different pig breeds have been captured from organised pig farms using cell phone camera. A pig breed dataset has been developed from those captured images. The dataset contains data for 250 individual pigs belonging to five breeds. Several well-established algorithms like SVM, DT, RF, LR, GBDT, NB, MLP and KNN have been carefully trained on the developed dataset and compared against each other based on AUC PR and AUC ROC metrics. Analysis shows that SVM with RBF kernel is the best algorithm for pig breeds classification with breed prediction accuracy of 98%. The authors are hopeful that the developed model can be used as ready to use technology to help recognising breed identity of a pig from its image.

**Figure 15** Confusion matrix for pig breed classification using SVM (see online version for colours)




Although our method predicts pig breeds very accurately, but some misclassifications occurred where two pigs belonging to two different breeds have very similar visual characteristics. In the future, we are planning to use deep learning-based approaches (Hoang et al., 2021; Thanh et al., 2022) to improve the accuracy of pig breed classification problem. We are planning to test different algorithms such as monarch butterfly optimisation (MBO) (Wang et al., 2019), earthworm optimisation algorithm (EWA) (Wang et al., 2018), elephant herding optimisation (EHO) (Wang et al., 2015), moth search (MS) algorithm (Wang, 2018), Slime mould algorithm (SMA) (Li et al., 2020a), hunger games search (HGS) (Yang et al., 2021), Runge Kutta optimiser (RUN) (Ahmadianfar et al., 2021), colony predation algorithm (CPA) (Tu et al., 2021), and Harris hawks optimisation (HHO) (Heidari et al., 2019) for pig breed prediction. We are also planning to include more pig breeds in the dataset so that classification can be more robust and can be used by more stakeholders.

## References

Ahmadianfar, I., Heidari, A.A., Gandomi, A.H., Chu, X. and Chen, H. (2021) 'RUN beyond the metaphor: an efficient optimization algorithm based on Runge Kutta method', *Expert Systems with Applic*ations, Vol. 181, p.115079, ISSN: 0957-4174, DOI: https://doi.org/10.1016/j.eswa.2021.115079 [online] https://www.sciencedirect.com/science/article/pii/S0957417421005200.

Andrew, W., Hannuna, S., Campbell, N. and Burghardt, T. (2016) 'Automatic individual holstein friesian cattle identification via selective local coat pattern matching in rgb-d imagery', in *IEEE International Conference on Image Processing (ICIP)*.

Arivazhagan, S., Shebiah, R.N., Ananthi, S. and Varthini, S.V. (2013) 'Detection of unhealthy region of plant leaves and classification of plant leaf diseases using texture features', *Agricultural Engineering International: CIGR Journal*, Vol. 15, No. 1, pp.211–217.

Armitage, D.W. and Ober, H.K. (2010) 'A comparison of supervised learning techniques in the classification of bat echolocation calls', *Ecological Informatics*, Vol. 5, No. 6, pp.465–473.

Ayanzadeh, A. and Vahidnia, S. (2018) *A Modified Deep Neural Networks for Dog Breeds Identification* [online] http://Preprints.org (accessed 24 June 2020).

Bollapragada, R., Byrd, R.H. and Nocedal, J. (2019) 'Exact and inexact subsampled Newton methods for optimization', *IMA Journal of Numerical Analysis*, Vol. 39, No. 2, pp.545–578.

Bottou, L. (2010) 'Large-scale machine learning with stochastic gradient descent', in *Proceedings of COMPSTAT*.

Burghardt, T. (2012) *A General Introduction to Visual Animal Biometrics*, Visual Information Laboratory.

Chan, T.F., Golub, G.H. and LeVeque, R.J. (1982) 'Updating formulae and a pairwise algorithm for computing sample variances', in *COMPSTAT 1982 5th Symposium held at Toulouse 1982*.

Chernov, V., Alander, J. and Bochko, V. (2015) 'Integer-based accurate conversion between RGB and HSV color spaces', *Computers & Electrical Engineering*, Vol. 46, pp.328–337, ISSN: 0045-7906, DOI: https://doi.org/10.1016/j.compeleceng.2015.08.005, https://www.sciencedirect.com/science/article/pii/S0045790615002827.

Chomboon, K., Chujai, P., Teerarassamee, P., Kerdprasop, K. and Kerdprasop, N. (2015) 'An empirical study of distance metrics for k-nearest neighbor algorithm', in *Proceedings of the 3rd International Conference on Industrial Application Engineering*.

Davis, J. and Goadrich, M. (2006) 'The relationship between precision-recall and ROC curves', in *Proceedings of the 23rd International Conference on Machine Learning*.

Fan, R-E., Chang, K-W., Hsieh, C-J., Wang, X-R. and Lin, C-J. (2008) 'LIBLINEAR: a library for large linear classification', *Journal of Machine Learning Research*, August, Vol. 9, pp.1871–1874.

Fawcett, T. (2006) 'An introduction to ROC analysis', *Pattern Recognition Letters*, Vol. 27, No. 8, pp.861–874.

Food and Agriculture Organization of the United Nations (2011) *Molecular Genetic Characterization of Animal Genetic Resources* [online] https://www.fao.org/3/i2413e/i2413e00.htm (accessed 24 June 2020).

Food and Agriculture Organization of the United Nations (2012) *Phenotypic Characterization of Animal Genetic Resources* [online] https://www.fao.org/3/i2686e/i2686e00.htm (accessed 24 June 2020).

Frasca, T.M., Sestito, A.G., Versek, C., Dow, D.E., Husowitz, B.C. and Derbinsky, N. (2016) 'A comparison of supervised learning algorithms for telerobotic control using electromyography signals', in *Thirtieth AAAI Conference on Artificial Intelligence*.

Friedman, J.H. (2002) 'Stochastic gradient boosting', *Computational Statistics & Data Analysis*, Vol. 38, No. 4, pp.367–378s.

Genoud, A.P., Gao, Y., Williams, G.M. and Thomas, B.P. (2020) 'A comparison of supervised machine learning algorithms for mosquito identification from backscattered optical signals', *Ecological Informatics*, Vol. 58, p.101090, ISSN: 1574-9541, DOI: https://doi.org/10.1016/j.ecoinf.2020.101090, https://www.sciencedirect.com/science/article/pii/S1574954120300406.

Gonzalez, R.C., Woods, R.E. and Eddins, S.L. (2003) *Digital Image Processing using MATLAB*, Prentice Hall, New Jersey.

Hailu, A. (2015) 'Breed characterization: tools and their applications', *Open Access Library Journal*, Vol. 2, No. 4, p.1.

Haralick, R.M., Shanmugam, K. and Dinstein, I.H. (1973) 'Textural features for image classification', *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, Vol. SMC-3, No. 6, pp.610–621, DOI: 10.1109/TSMC.1973.4309314.

He, K. and Sun, J. (2015) 'Convolutional neural networks at constrained time cost', in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Heidari, A.A., Mirjalili, S., Faris, H., Aljarah, I., Mafarja, M. and Chen, H. (2019) 'Harris hawks optimization: algorithm and applications', *Future Generation Computer Systems*, Vol. 97, pp.849–872, ISSN: 0167-739X, DOI: https://doi.org/10.1016/j.future.2019.02.028, https://www.sciencedirect.com/science/article/pii/S0167739X18313530.

Hoang, T.D.N., Thi, T.L., Ugur, E., Aditya, K. and Surya, P.V. (2021) 'Dermoscopic image segmentation method based on convolutional neural networks', *International Journal of Computer Applications in Technology*, Vol. 66, No. 2, pp.89–99.

Hoffmann, I. and Scherf, B. (2010) 'Implementing the global plan of action for animal genetic resources', *Animal Genetic Resources*, Vol. 47, pp.1–10, DOI: 10.1017/S2078633610001050.

Jwade, S.A., Guzzomi, A. and Mian, A. (2019) 'On farm automatic sheep breed classification using deep learning', *Computers and Electronics in Agriculture*, Vol. 167, pp.150–155, DOI: https://doi.org/10.1016/j.compag.2019.105055, https://www.sciencedirect.com/science/article/pii/S0168169918318465.

Karlik, B. and Olgac, A.V. (2011) 'Performance analysis of various activation functions in generalized MLP architectures of neural networks', *International Journal of Artificial Intelligence and Expert Systems*, Vol. 1, No. 4, pp.111–122.

Kingma, D.P. and Ba, J. (2014) *Adam: A Method for Stochastic Optimization*, arXiv preprint arXiv:1412.6980.

Kleinbaum, D.G., Dietz, K., Gail, M., Klein, M. and Klein, M. (2002) *Logistic Regression: A Self-Learning Text*, pp.1–39, Springer, New York, NY, ISBN: 978-1-4419-1742-3, DOI: 10.1007/978-1-4419-1742-3_1, https://doi.org/10.1007/978-1-4419-1742-3_1.

Kukreja, M., Johnston, S.A. and Stafford, P. (2012) 'Comparative study of classification algorithms for immunosignaturing data', *BMC Bioinformatics*, Vol. 13, No. 1, p.139.

Kumar, S., Pandey, A., Satwik, K.S.R., Kumar, S., Singh, S.K., Singh, A.K. and Mohan, A. (2018) 'Deep learning framework for recognition of cattle using muzzle point image pattern', *Measurement*, Vol. 116, pp.1–17, ISSN: 0263-2241, DOI: https://doi.org/10.1016/j.measurement.2017.10.064, https://www.sciencedirect.com/science/article/pii/S0263224117306991.

Lahiri, M., Tantipathananandh, C., Warungu, R., Rubenstein, D.I. and Berger-Wolf, T.Y. (2011) 'Biometric animal databases from field photographs: identification of individual zebra in the wild', in *Proceedings of the 1st ACM International Conference on Multimedia Retrieval*.

Li, G., Wang, G-G., Dong, J., Yeh, W-C. and Li, K. (2021a) 'DLEA: a dynamic learning evolution algorithm for many-objective optimization', *Information Sciences*, Vol. 574, pp.567–589, ISSN: 0020-0255, DOI: https://doi.org/10.1016/j.ins.2021.05.064, https://www.sciencedirect.com/science/article/pii/S0020025521005363.

Li, S., Chen, H., Wang, M., Heidari, A.A. and Mirjalili, S. (2020a) 'Slime mould algorithm: a new method for stochastic optimization', *Future Generation Computer Systems*, Vol. 111, pp.300–323, ISSN: 0167-739X, DOI: https://doi.org/10.1016/j.future.2020.03.055, https://www.sciencedirect.com/science/article/pii/S0167739X19320941.

Li, W., Wang, G.G. and Alavi, A.H. (2020b) 'Learning-based elephant herding optimization algorithm for solving numerical optimization problems', *Knowledge-Based Systems*, Vol. 195, p.105675, ISSN: 0950-7051, DOI: https://doi.org/10.1016/j.knosys.2020.105675, https://www.sciencedirect.com/science/article/pii/S0950705120301179.

Li, W., Wang, G.G. and Gandomi, A.H. (2021b) 'A survey of learning-based intelligent optimization algorithms', *Archives of Computational Methods in Engineering*, 1 August, Vol. 28, No. 5, pp.3781–3799, ISSN: 1886-1784, DOI: 10.1007/s11831-021-09562-1, https://doi.org/10.1007/s11831-021-09562-1.

Liaw, A. and Wiener, M. (2002) 'Classification and regression by random forest', *R News*, Vol. 2, No. 3, pp.18–22.

Liu, D.C. and Nocedal, J. (1989) 'On the limited memory BFGS method for large scale optimization', *Mathematical Programming*, Vol. 45, Nos. 1–3, pp.503–528.

Liu, X., Xia, T., Wang, J. and Lin, Y. (2016) *Fully Convolutional Attention Localization Networks: Efficient Attention Localization for Fine-Grained Recognition*, arXiv preprint arXiv: 1603.06765, Vol. 1, No. 2, p.4.

Loh, W-Y. (2011) 'Classification and regression trees', *WIREs Data Mining and Knowledge Discovery*, Vol. 1, No. 1, pp.14–23, DOI: https://doi.org/10.1002/widm.8 [online] https://wires.onlinelibrary.wiley.com/doi/abs/10.1002/widm.8; eprint = https://wires.onlinelibrary.wiley.com/doi/pdf/10.1002/widm.8.

Mandal, S.N., Ghosh, P., Mukherjee, K., Dan, S., Mustafi, S., Roy, K., Hajra, D.K. and Banik, S. (2020) 'InceptGI: a ConvNet-Based classification model for identifying goat breeds in India', *Journal of The Institution of Engineers (India): Series B*, 1 October, Vol. 101, No. 5, pp.573–584, ISSN: 2250-2114, DOI: 10.1007/s40031-020-00471-8, https://doi.org/10.1007/s40031-020-00471-8.

Meena, S.D. and Agilandeeswari, L. (2019) 'An efficient framework for animal breeds classification using semi-supervised learning and multi-part convolutional neural network (MP-CNN)', *IEEE Access*, Vol. 7, pp.151783–151802, DOI: 10.1109/ACCESS.2019.2947717.

Ministry of Fisheries, Animal Husbandry & Dairying (2019) *20th Livestock Census*, Animal Husbandry Statistics Division, Government of India, New Delhi [online] https://pib.gov.in/PressReleasePage.aspx?PRID=1588304 (accessed 24 June 2020).

Osisanwo, F.A.A.J., Awodele, O., Hinmikaiye, J., Olakanmi, O. and Akinjobi, J. (2017) 'Supervised machine learning algorithms: classification and comparison', *International Journal of Computer Trends and Technology (IJCTT)*, Vol. 48, No. 3, pp.128–138.

Parkhi, O.M., Vedaldi, A., Zisserman, A. and Jawahar, C.V. (2012) 'Cats and dogs', in *IEEE Conference on Computer Vision and Pattern Recognition*.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R. and Dubourg, V. (2011) 'Scikit-learn: machine learning in Python', *Journal of Machine Learning Research*, Vol. 12, pp.2825–2830.

Peng, W., Yang, H., Cai, K., Zhou, L., Tan, Z. and Wu, K. (2019) 'Molecular identification of the Danzhou chicken breed in China using DNA barcoding', *Mitochondrial DNA Part B*, Vol. 4, No. 2, pp.2459–2463.

Pica, G., Pica-Ciamarra, U. and Otte, J. (2008) 'The livestock sector in the World Development Report 2008: re-assessing the policy priorities', *PPLPI Research Report*.

Pilario, K.E., Shafiee, M., Cao, Y., Lao, L. and Yang, S-H. (2020) 'A review of kernel methods for feature extraction in nonlinear process monitoring', *Processes*, Vol. 8, No. 1, Article No. 24, ISSN: 2227-9717, DOI: 10.3390/pr8010024 [online] https://www.mdpi.com/2227-9717/8/1/24.

Hearst, M.A., Dumais, S.T., Osuna, E., Platt, J. and Scholkopf, B. (1998) 'Support vector machines', *IEEE Intelligent Systems and their Applications*, Vol. 13, No. 4, pp.18–28, DOI: 10.1109/5254.708428.

Raschka, S. (2018) *Model Evaluation, Model Selection, and Algorithm Selection in Machine Learning*, arXiv preprint arXiv:1811.12808.

Rish, I. (2001) 'An empirical study of the naive Bayes classifier', in *IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence*.

Ruck, D.W., Rogers, K. and Kabrisky, M. (1990) 'Feature selection using a multilayer perceptron', *Journal of Neural Network Computing*, Vol. 2, No. 2, pp.40–48.

Sardina, M.T., Tortorici, L., Mastrangelo, S., Di Gerlando, R., Tolone, M. and Portolano, B. (2015) 'Application of microsatellite markers as potential tools for traceability of Girgentana goat breed dairy products', *Food Research International*, Vol. 74, pp.115–122, ISSN: 0963-9969, DOI: https://doi.org/10.1016/j.foodres.2015.04.038, https://www.sciencedirect.com/science/article/pii/S096399691500201X.

Sheshasaayee, A. and Thailambal, G. (2017) 'Comparison of classification algorithms in text mining', *International Journal of Pure and Applied Mathematics*, Vol. 116, No. 22, pp.425–433.

Sundaram, D. and Loganathan, A. (2020) 'A new supervised clustering framework using multi discriminative parts and expectation-maximization approach for a fine-grained animal breed classification (SC-MPEM)', *Neural Processing Letters*, 1 August, Vol. 52, No. 1, pp.727–766, ISSN: 1573-773X, DOI: 10.1007/s11063-020-10246-3, https://doi.org/10.1007/s11063-020-10246-3.

Thanh, T.T., Hoang, T.D.N. and Quoc, H.N. (2022) 'A dish recognition framework using transfer learning', *IEEE Access*, Vol. 10, pp.7793–7799, DOI: 10.1109/ACCESS.2022.3143119.

Tu, J., Chen, H., Wang, M. and Gandomi, A.H. (2021) 'The colony predation algorithm', *Journal of Bionic Engineering*, 1 May, Vol. 18, No. 3, pp.674–710, ISSN: 2543-2141, DOI: 10.1007/s42235-021-0050-y, https://doi.org/10.1007/s42235-021-0050-y.

Wang, G-G. (2018) 'Moth search algorithm: a bio-inspired metaheuristic algorithm for global optimization problems', *Memetic Computing*, 1 June, Vol. 10, No. 2, pp.151–164, ISSN: 1865-9292, DOI: 10.1007/s12293-016-0212-3, https://doi.org/10.1007/s12293-016-0212-3.

Wang, G-G., Deb, S. and Coelho, L.D.S. (2015) 'Elephant herding optimization', in *3rd International Symposium on Computational and Business Intelligence (ISCBI).*

Wang, G-G., Deb, S. and Coelho, L.D.S. (2018) 'Earthworm optimisation algorithm: a bio-inspired metaheuristic algorithm for global optimisation problems', *International Journal of Bio-Inspired Computation*, Vol. 12, No. 1, pp.1–22.

Wang, G-G., Deb, S. and Cui, Z. (2019) 'Monarch butterfly optimization', *Neural Computing and Applications*, 1 July, Vol. 31, No. 7, pp.1995–2014, ISSN: 1433-3058, DOI: 10.1007/s00521-015-1923-y, https://doi.org/10.1007/s00521-015-1923-y.

Wang, X., Liu, J., Zhou, G., Guo, J., Yan, H., Niu, Y., Li, Y., Yuan, C., Geng, R. and Lan, X. (2016) 'Whole-genome sequencing of eight goat populations for the detection of selection signatures underlying production and adaptive traits', *Scientific Reports*, 12 December, Vol. 6, No. 1, p.38932, ISSN: 2045-2322, DOI: 10.1038/srep38932, https://doi.org/10.1038/srep38932.

Yang, Y., Chen, H., Heidari, A.A. and Gandomi, A.H. (2021) 'Hunger games search: visions, conception, implementation, deep analysis, perspectives, and towards performance shifts', *Expert Systems with Applications*, Vol. 177, p.114864, ISSN: 0957-4174, DOI: https://doi.org/10.1016/j.eswa.2021.114864, https://www.sciencedirect.com/science/article/pii/S0957417421003055.

Yu, X., Wang, J., Kays, R., Jansen, P.A., Wang, T. and Huang, T. (2013) 'Automated identification of animal species in camera trap images', *EURASIP Journal on Image and Video Processing*, 4 September, Vol. 2013, No. 1, p.52, ISSN: 1687-5281, DOI: 10.1186/1687-5281-2013-52, https://doi.org/10.1186/1687-5281-2013-52.

Zhang, T., Wiliem, A., Hemsony, G. and Lovell, B.C. (2015) 'Detecting kangaroos in the wild: the first step towards automated animal surveillance', in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).*

## Abbreviations

| | |
|---|---|
| SVM | Support vector machine |
| LR | Logistic regression |
| NB | Naïve Bayes |
| MLP | Multilayer perceptron |
| DT | Decision tree |
| RF | Random forest |
| KNN | K-nearest neighbours |
| GBDT | Gradient boosted decision trees |
| AUC | Area under curve |
| PR | Precision-recall |
| ROC | Receiver operating characteristic. |