Graph based Classification Techniques for Pig Breed Identification from Hand-crafted Visual Muzzle Descriptors

A

Thesis submitted

for the award of the degree of

DOCTOR OF PHILOSOPHY

By

Shoubhik Chakraborty

DEPARTMENT OF ELECTRONICS AND ELECTRICAL ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY GUWAHATI

GUWAHATI - 781 039, ASSAM, INDIA

AUGUST 2022

Certificate

This is to certify that the thesis entitled "**Graph based Classification Techniques for Pig Breed Identification from Hand-crafted Visual Muzzle Descriptors**", submitted by **Shoubhik Chakraborty** (166102007), a research scholar in the *Department of Electronics and Electrical Engineering, Indian Institute of Technology Guwahati*, for the award of the degree of **Doctor of Philosophy**, is a record of an original research work carried out by him under my supervision and guidance. The thesis has fulfilled all requirements as per the regulations of the institute and in my opinion has reached the standard needed for submission. The results embodied in this thesis have not been submitted to any other University or Institute for the award of any degree or diploma.

Dated:

Guwahati.

Dr. Kannan Karthik
Associate Professor
Dept. of Electronics and Electrical Engg.
Indian Institute of Technology Guwahati
Guwahati - 781 039, Assam, India.

To

The Almighty

My beloved parents

for their blessings, love and encouragement

My brother **Saurav**

for his moral support and encouragement

My wife **Arundhati**

for her love and sacrifice

&

My guide **Dr. Kannan Karthik**

for his guidance and inspiration

Acknowledgements

I am obliged to GOD for His divine guidance and blessings. I solely dedicate my thesis to Lord Shri Krishna.

This thesis would not have been possible without the immense help and support of several people in various measures. I would like to convey my acknowledgment to all of them.

First and foremost, I express my sincere gratitude to my research supervisor, Dr. Kannan Karthik for providing me an opportunity to work under his guidance. It is very difficult to describe my feelings in words to acknowledge my supervisor for his continuous guidance in all aspects, constant motivation and support throughout the doctoral studies. I am very much thankful to him for transforming me from an unstructured form to a structured form in every aspect of my life and showing me a different path of life. It would be completely impossible for me to bring the research as well as the thesis to this form without the immense facilities provided by him in the Signal Processing Laboratory and the freedom of work he has given to me.

I am thankful to my doctoral committee members Prof. M. K. Bhuyan, Dr. Suresh Sundaram and Dr. V. Vijaya Saradhi for their encouragement and valuable suggestions on my work.

I am also thankful to Dr. Santanu Banik for his valuable suggestions related to the biological aspects of my research work which helped me a lot in the understanding of the problem.

I would like to thank faculty members and the office staffs of the Department of Electronics and Electrical Engineering, IIT Guwahati, for their help in carrying out this research work. I am very much thankful to Mr. Mukut Sharma and Mr. Dasarath Das and Mr. Sundeep Borah for their kind support and help.

I am thankful to my friends Sandeep, Sibasis, Dr. Deepika, Shikha, Mrinmoy, Dr. Protima, Dr. Sisir, Dr. Vikram, Dr. Akhilesh, Balaji, Moa, Saswati, Anik and other members of the Signal Processing Lab for their help and support during my research work.

I am thankful to my parents, brother and wife for their sacrifice and support. Without their

constant support and encouragement it would not have been possible for me to complete the thesis work.

Shoubhik Chakraborty

Abstract

Non-intrusive and automated detection of pig breeds, particularly from a machine-vision standpoint, is important from a commercial perspective for both vendors and buyers. The muzzle of a pig as a nasal interface, constitutes a robust biometric descriptor, whose morphological features when agglomerated, can be used to isolate specific breeds.

In this work, hand-crafted color and texture descriptors/statistics have been selected for profiling four pig-breeds: Duroc, Ghungroo, Hampshire and Yorkshire. While these hand-crafted visual descriptors by themselves are fairly robust and discriminative, it was recognized that classification accuracies can be improved, by controlling the decision space either by choosing the feature-type based on colour or texture or by combining features and also selecting the order in which particular breeds are siphoned.

In that light, in the first part of this work, a stable, hierarchical breed-siphoning policy is proposed, where breed-types are identified and distilled in specific serial order. While the existing Phylogenetic Hierarchical Agglomerative Clustering algorithm (AGNES) also produces a siphoning hierarchy, the tree is different from the proposed one on one main front: The AGNES-tree feeds on a general minimum distance table evaluated across different feature sets (thus is feature-type agnostic and therefore not optimized); On the other hand, the proposed tree feeds on secondary, feature specific cumulative distances derived from the primary cluster-distance table to identify which feature-type can be used at each decision node in the breed-siphoning tree. This optimizes the siphoning procedure and makes clusters at each level more compact ensuring better use of linear SVM models. The accuracies reported for AGNES versus proposed-siphoning tree were: (83.58% vs 86.45%) for Duroc, (94.11% vs 93.02%) for Ghungroo, (81.78% vs 86.91%) for Hampshire and (96.19% vs 98.54%) for Yorkshire.

To improve the accuracies further, two major problems which had to be confronted, while handling muzzle datasets were (i) limited training data (5-7 pigs per breed only) and (ii) High variability from the point of view of colour and texture within each breed class, particularly for breeds such as Duroc and Hampshire. This meant that the isometric assumption made by Linear SVM models would prove sub-optimal (leading to an over-generalization of the classes). Hence,

a data-dependent and data-associative model, graphically inclined, would have to be synthesized from the training breed-samples. In this context a Maximum Spanning Tree (MaxST) algorithm has been proposed, to first characterize the breed classes. Once the MaxST model is built from the training nodes of a given class/breed, an inductive procedure with respect to the test-sample (being checked) is deployed to create a new tree and then a check is performed to see if the primary MaxST structure connecting the training nodes from the given class/breed have been altered. The larger the change in the primary structure, the higher is the outlier score assigned to the test point with respect to the given class/breed.

Furthermore, it was observed that with this data-associative tree arrangement it was possible to resolve partially overlapping clusters (the inlier-inlier problem) and to large extent trap the shape of the individual cluster. Ownership resolution in the case of the inlier-inlier arrangement was done in a probabilistic light by deploying a sequence of independent random projection matrices, to reduce the feature space to a sequence of randomly oriented 2D-planes. To quantify the inlier-degree (or the depth of the point within the projected cluster), a cumulative scoring approach based on triangles formed between several training nodes was deployed. The breed-class which registered the lowest outlier score was declared the owner of the test point. Deep inliers were expected to register very low scores and borderline inliers close to as high as 50%, with respect to a particular breed-model.

However, this method of resolving the inlier-inlier conflict was computationally expensive. While, the outlier score generated based on the MaxST formulation proved to be sensitive to the proximity of the test point to the training cluster, when outside the cluster; but when present within the cluster, a zero differential score was registered. In such a case, the dual of the MaxST, i.e. a Minimum Spanning Tree(MinST) was found to be useful. In contrast to the formulation based on MaxST, the outlier score generated based on an equivalent MinST formulation was sensitive to the positional arrangement of the test point lying within the cluster. Thus the MaxST formulation in conjunction with its MinST counterpart is a robust and computationally efficient outlier detection framework. In a multi-class classification scenario using this outlier detection framework, the test point is assigned the class label for which it has the lowest outlier

score.

Finally, the outlier-check based frame aided by the cumulative scoring approach based on triangles formed between several training nodes was extended to a multi-class setting, with some minor feature adaptation and this showed a considerable improvement in accuracy in relation to the earlier siphoning proposition. When completely different sets of pigs were used for training and testing (50-50 split), the proposed algorithm reported relatively high mean classification accuracies of 93.26% for Duroc, 97.19% for Ghungroo, 93.61% for Hampshire and 100% for Yorkshire. Similar classification accuracies were achieved with the MaxST-MinST outlier framework as well but with a much lesser computational cost. The highlight was the $5 - 7\%$ increase in accuracy for the difficult breeds, Duroc and Hampshire with the data-associative spanning tree based outlier detection frame.

Keywords: Pig Breeds, Segmentation, Gradient Significance Map, Morphological Tophat operator, Hierarchical Classification Scheme, Outlier Detection, Maximum Spanning Tree, Minimum Spanning Tree.

Contents

List of Figures

List of Tables

1

Introduction

Contents

1.1 Introduction

Biometric markers are used extensively for human subjects, for individual or ethnic identification, for surveillance, classification etc. They are being increasingly applied to animals, again either to track an individual animal or a breed. An example of individual animal biometric application is tracking a champion race horse for its immediate value and in future for breeding. On the other hand, a typical breed identification application could be for animals such as cows, goats and pigs, whose "produce", in some form or the other is consumed (i.e. has commercial value and impact). Here, the identity of the cow or pig in a larger group context, i.e. its breed-type, becomes crucial. For instance, it has been found and recorded that the following four specific breeds of cows viz. Gir cow of Gujarat; Rathi cow of states Uttar Pradesh, Madhya Pradesh and Haryana; Red Sindhi cow of states Punjab, Haryana, Karnataka; Sahiwal cow of states Uttar Pradesh, Haryana and Madhya Pradesh, can produce large quantities of nutritious milk every day [3]. In the same vein, in the context of pigs, most farms prefer pure-breeds as opposed to cross-breeds from the point of view of reproduction and pork quality.

Aspects connected to their natural health condition in current local environments, weight and size associated with specific pig-breeds, play a crucial role in formulating a preference for certain breed-types over others. The Large White Yorkshire, Hampshire, Duroc etc., are among the few imported exotic breeds in India [4]. The Yorkshire breed for instance is popular because it lends itself to cross-breeding. Being a prolific breeder, it provides a good amount of meat for consumption. Hampshire hogs on the other hand are noted for being well-muscled and rapid growers, and for exhibiting good carcass quality when used as meat animals. Further, this breed is also used extensively in different cross-breeding programmes due to its preferential black coat colour in the local market. The Duroc breed is noted for its excellent weight gain and feed efficiency, which simplifies maintenance protocols considerably. This breed is mostly used as terminal sire in cross-breeding for lean meat production.

Most biometrics are found concentrated near sensory interfaces e.g. muzzle of pig [5] or facial profile in cows [6] or hoof patterns in horses [7]. Aggregates of these biometric traces at a

macro level, may qualify as what are known as "breed identifiers". For instance, the face, which is a crucial biometric from a social standpoint of human beings, can be analyzed at a coarser level involving colour, texture, facial feature skewness/bias, to segregate people hailing from different types of ethnic communities. These facial parameters are transmitted from generation to generation, as long as the ethnic group stays insular. In a similar fashion in the case of pigs, for a particular set of morphological parameters to qualify as a breed indicator the following conditions must be satisfied:

- There must be a biometric connection, i.e. these morphological parameters must be linked to their ancestors.
- Furthermore, these parameters should be concentrated near sensory interfaces [8], at places where there is a continuous interaction with the surrounding environment. Over generations some of these macro-morphological parameters will stabilize and passed down to future generations.

However, when it comes to the practical deployment of a breed profiling procedure, it is important to first identify the domain (visual, acoustic, chemical etc.), and then the nature of the measurements taken in that domain, which may qualify as breed-relevant features or statistics. Upon entering the digital/analog space of measurements, several challenges may arise:

- Among the gamut of measurements which have been hypothesized, which ones are the best choices and why? This, "why", is not easy to answer, particularly when one does not have sufficient on-field data (both in terms of time-trends, type and numbers).
- Taking measurements from a particular animal, is itself a difficult task, depending on how intrusive the process is. Unlike a human subject, who co-operates while gathering his/her biometric such as face, fingerprint or iris scan, animal subjects need to be restrained in most cases to freeze the animal's natural interaction with the external environment, before taking the measurement. Some biometric-measurements such as animal-gait [9], animal-sounds [10], breathing-sounds [11], sniffing-patterns [12] are easier to observe and gather

from a distance as compared to others such as facial-signatures [6], chemical composition of excreta [13].

- Once the mode of acquisition and type of measurement is established, it is to be checked whether these measurements remain stable with time and are distinct for different breeds. This is the most important form of analysis which is connected with the signal processing and analysis domain. It has been noted that pigs beyond the weaning period (more than six weeks since birth), tend to exhibit stability in health and other parameters [14].

It is first important to establish the notion of a breed identifier and then over successive iterations, water it down to a handful of measurements or statistics in some domain (visual or acoustic or chemical etc.), so that finally there is a mechanism for easily deriving a handful of numbers based on some form of sensory data which can be be quickly used to establish the breed of the animal.

In this work, the focus is on breed identifiers associated with four breeds of pigs viz. Duroc, Ghungroo, Hampshire and Yorkshire. While Duroc, Hampshire and Yorkshire are imported breeds in India, Ghungroo is a native breed found in the regions of north Bengal. Visual identifiers and descriptors such as those involving still images of faces of pigs or parts of the face, such as the tip of the snout (termed as the muzzle), have been found to be stable yet discriminative across breeds [5], when post-processed, signal-conditioned and harvested carefully.

1.2 Biometric and Breed Identifiers in Pigs

It was surmised in [5], that biometric identifiers tend to be concentrated near sensory interfaces. These sensory interfaces could be the eye (ocular type with the biometric in question being the iris) [15], the ear (aural type dealing with vein patterns on the inner surface) [16], the nose or snout (nasal and tactile type especially for pigs, involving the muzzle) [17, 18] and many others. While a single biometric identifier imparts a distinct identity to a particular pig, it does not directly qualify the pig in a larger group (analogous to "race" based classification for humans based on their origin and life-style over generations). When several biometric micro-

identifiers are aggregated, at a coarser level they constitute a "breed-identifier", which qualifies the pig in a larger setting based on some similarities in traits.

1.3 Muzzle as a Pig-breed Identifier

The nasal disc on the snout of the pig, while rigid enough to be used for digging, has numerous sensory receptors. Pigs use their snout while exploring and searching for food items, to push objects, to flatten them, for scooping and for leveraging out thick roots. Furthermore, under natural conditions, pigs may spend 75% of their daily activity engaged in rooting and foraging [19]. Apart from this, snout also houses the nasal interface carrying two nostrils as shown in Fig. 1.1, which plays a crucial role on several fronts such as: (i) breathing, (ii) discriminating between odors and tracking down food, (iii) using the odors stemming from chemical discharges to locate one's mate [19]. This frontal portion or disk is termed as the "muzzle". Thus, the muzzle segment serves as both a tactile as well as a nasal interface for interacting with the environment. Owing to the presence of many sensory features in virtually the same location, it was concluded in our earlier work [5], that the frontal image of the muzzle covering the heart shaped nasal disc which houses two nostrils could be used as a "breed identifier".

The pigs selected for analysis and profiling have crossed what is known as the weaning period (more than six weeks from the date of birth). This implies that their biometrics and breed-identities have crystallized morphologically in time. Beyond this weaning period virtually no change is observed in the biometric and breed-parameters [20].

The manually cropped muzzle segment of a Hampshire pig is shown in Fig. 1.1 in which the key parts of a muzzle are marked. Apart from having two large nostrils and hair at the periphery of the muzzle contour, the muzzle region also shows large sweat pores and some stunted hair. This stunted hair was restricted to the darker region leaving the pinkish-white patch barren with respect to the hair. Sweat pores, however, were present uniformly in both the dark and pinkish white patches.

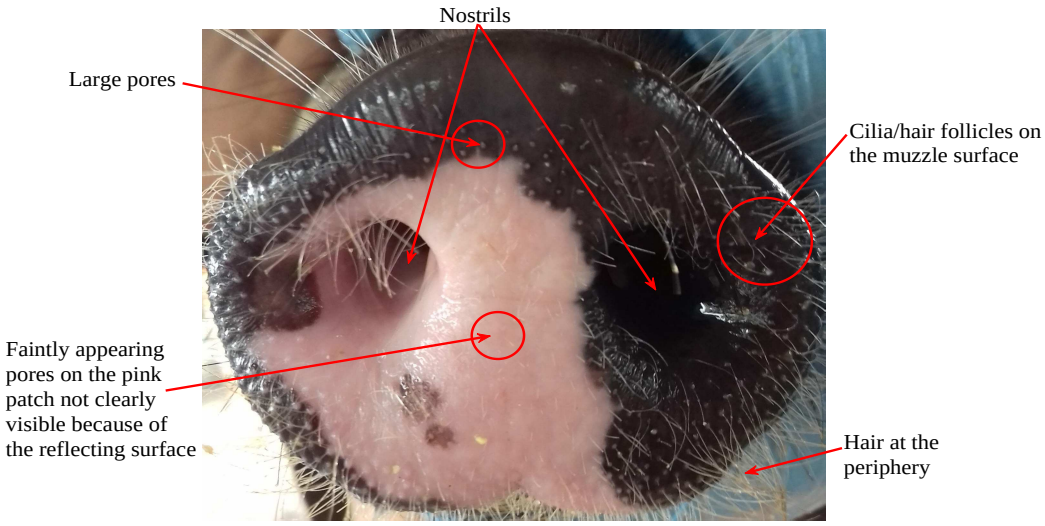


Figure 1.1: Description of the muzzle of a Hampshire pig.

1.4 Visual Descriptors Related to Muzzle Surface for Breed Identification

The visual domain provides ample information for breed discrimination at a machine level. The cropped frontal view of the muzzle of the pig was chosen as the input muzzle image for breed model building and analysis. Exemplar muzzle images from each of the four breeds viz. Duroc, Ghungroo, Hampshire and Yorkshire are shown in Fig. 1.2. On a breed specific basis, the following visual trends were observed for the four breeds:

- Ghungroo: These pigs have a high density of sweat-pores and cilia (hair follicles), more towards the periphery of the muzzle (near the muzzle contour) and towards the bottom side of the muzzle surface below the nostrils. Muzzle is either completely black or blackish-grey. Specular artefacts are produced upon imaging due to nasal secretions.

- Duroc: These pigs either have a purely powdery black-coloured muzzle or are dual coloured with a relatively small pink region embedded within the black area. Sweat pores are not as prominent as Ghungroo and limited amount of cilia present over the muzzle surface. Also the muzzle contour of the Duroc pigs contain two notches at approximate angular location of 60° and 120° in the upper half of the contour.

(a) Duroc (b) Ghungroo (c) Hampshire (d) Yorkshire

Figure 1.2: Exemplar muzzle images from the four breeds. Each column contains two examples from the same breed.

- Yorkshire: These pigs have a completely pink coloured muzzle surface with almost no hair or cilia. Sweat pores are present, but not visually prominent.
- Hampshire: They have a dual coloured muzzle surface (greyish-black and pink) with a partial pink patch. All male Hampshire pigs exhibit this dual colouration. Pink patch has an arbitrary size and orientation. The sweat pores and cilia are concentrated more over the greyish part of the muzzle. The pink region is devoid of any cilia and very less prominent sweat pores.

The visual information listed above in a breed specific manner indicate that the muzzle images from the four breeds have some distinctive features in certain information domains viz. muzzle contour shape, colour and texture density on the muzzle surface. Discriminative visual descriptors can be extracted from these domains to segregate the breeds based on their muzzle image. These information domains are described next:

- Texture density: These refer to the spatial distribution of lines/curves (from cilia/hair follicles on the muzzle surface), small dots of different sizes (from sweat pores on the muzzle surface) and in some cases internal contours and curves (from the boundary of

pink regions and nostrils on the muzzle surface). While internal contours resulting from the boundary of pink region in dual coloured muzzle images can be used for individual biometric identification, they cannot be used as breed identifiers.

- Colour: As already mentioned, on the colour front Ghungroo muzzle images are purely greyish-black, Yorkshire muzzle images are purely pink, Hampshire muzzle images are dual coloured (grey and pink), while Duroc muzzle images are either dual-coloured or purely powdery black. For the Duroc, the fraction of pink region (if present) within the muzzle area is very small; for most cases it is less than 10%, while in very few cases it is around $20 - 25\%$. However, in Hampshire pigs, this fraction can be quite large reaching upto around 75% in some cases and the variation is also large with some pigs having the fraction of pink region around 25%. Because of the muzzle surfaces being specular in nature and due to improper lighting conditions, the colour profile of the muzzle images suffers from distortion. This leads to overlap of feature vectors capturing the colour information from different breeds in feature space, especially between Duroc and Hampshire.

- Shape: The muzzle contour shapes of Hampshire and Yorkshire are similar and takes the form of a inverse cardioid. Ghungroo muzzle muzzle contours contours are largely circular. Duroc has a distinct shape with two upper contour notches as compared to the other three breeds.

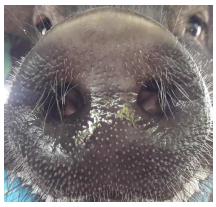
1.5 Stable Information Domains

The discussions in the previous section qualify the Texture density, Colour and Shape as robust information domains for breed classification. However, it is imperative that these domains of information be inspected from the viewpoint of practical implementation as well before they can be really churned to extract robust descriptors for breed classification. In order to construct accurate muzzle shape descriptors, it is imperative that the muzzle contour be extracted out with utmost accuracy which in itself is an extremely difficult task. It will be shown in

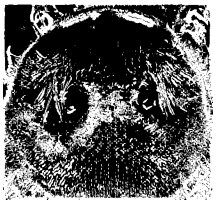
Chapter 2 that even with state-of-the-art active contour methods it is very difficult to extract out the muzzle contour precisely. In contrast there are a plethora of operators available in the literature based on gray level intensity difference, spatial distribution of gray levels etc. to pick up the texture information of the muzzle surface. Also the colour information of the muzzle surface is directly available from the [R G B] colour vector of each pixel. Hence, in this work instead of going for shape domain, the focus is on extracting robust visual descriptors from the texture and colour domain.

1.5.1 Texture Descriptors

The role of a texture descriptor is to highlight out the textural details like cilia, sweat pores etc. on the muzzle surface. There are several texture descriptors in literature [21]. The textural details on the muzzle surface like cilia and pores do not have any breed-specific correlation in their positional association. It is the amount of textural details that is of interest for breed segregation and these details are to be highlighted at their respective location on the muzzle surface with the help of a suitable filtering operator. Two different texture filters have been used to amplify the relevant breed-specific textural details while suppressing the other details as noise: one the Gradient Significance Map (GSM) [5,17] and the other uses the Morphological Tophat transformation operator [1]. As will be discussed in detail in Chapter 3, the response of the muzzle image to a filter operator is thresholded to obtain a binary map, which highlights the relevant textural details. When a muzzle image belonging to a pig-breed is passed through a texture filter, it is expected to enhance and amplify the features linked to that breed. For instance, the binary map obtained after applying the GSM tends to bring out the hair and pore patterns along with the contour-profiles, if prominent, in a particular breed(example Ghungroo). On the other hand, it suppresses the less prominent sweat pore patterns seen in Yorkshire and thus, in the processed image, the muzzle interior is found to be largely empty as shown in Fig. 1.3.



(a) Ghungroo Muzzle Image



(b) Binary Map
Ghungroo Muzzle Image

(c) Yorkshire Muzzle Image

(d) Binary Map
Yorkshire Muzzle Image

Figure 1.3: Binary maps generated using the GSM operator for Ghungroo and Yorkshire muzzle image. The cilia and large pores in the Ghungroo get highlighted whereas the smaller pores in Yorkshire are not brought out in the binary map.

1.5.2 Colour Descriptors

The role of a colour descriptor is to bring out selective attributes from the colour profile of the muzzle image which can impart distinctiveness to the breeds. One major impediment in the extraction of robust colour descriptors comes from improper illumination conditions while capturing snapshots of the muzzle images. Hence, the luminance information is to be separated out from the chromatic part before any further analysis. This is done by analysing the colour profile of the muzzle image in the $C_b - C_r$ space after transforming it from RGB domain to YCbCr domain. 2D-histogram is formed from the colour profile of the muzzle image in the $C_b - C_r$ domain. Breed-specific statistics are extracted out from this histogram in the following way:

- Using moments and moment linked statistics to trap certain aspects linked to the histogram shape such as bi-modality via Sarle's index [22] etc. This helps in isolating breeds such as Hampshire from the rest.

- Position, size and skew linked to the footprint of the colour histogram can be done via an Eigen-analysis in the $C_b - C_r$ space.

Table 1.1: Table listing the nature of the $C_b - C_r$ histogram for the different breeds.

Breed-type	$C_b - C_r$ **- histogram position/mean**	$C_b - C_r$ **- histogram footprint size**	$C_b - C_r$ **- histogram footprint skew**
Duroc	Close to origin as colour is black except for dual coloured muzzle images.	Small except for dual coloured muzzle images.	Minimal. Largely isometric except for dual coloured muzzle images.
Ghungroo	Close to origin as colour is black.	Mostly small.	Minimal and largely isometric.
Hampshire	Histogram is bimodal and positioned with one foot in the third quadrant (pinkish-red tinge) and the other near the origin(partially black).	Large	Slightly unpredictable but largely elongated and narrow.
Yorkshire	Histogram is in the third quadrant alone (purely pinkish-red tinge)	Small	Minimal and largely isometric.

Table 1.1 shows the nature of the $C_b - C_r$ histogram in terms of its location(which is equivalent to the centroid or mean), size of the histogram footprint and the skew associated with this footprint as a function of the four breeds. While the $C_b - C_r$ histogram positions are largely distinct for each of the four breeds, the size and the skew associated with the footprint of this histogram impart a distinctive feature to the $C_b - C_r$ histogram of Hampshire muzzle images.

1.6 Classification frames

Once robust feature sets have been constructed from the colour and texture descriptors, suitable classification frames need to be explored based on the nature of the feature sets. Two different classification frames have been used, which are discussed next:

Table 1.2: Table listing the different breed pairs and the feature sets which best separates the pair.

Breed Pair	Reliable Feature Type/ Combination	Justification
Duroc vs Ghungroo	Texture Alone	Colour profiles are similar (both have pre-dominantly dark muzzle).
Duroc vs Hampshire	Texture Alone	Duroc has moderate texture and is largely dark coloured (a few dual coloured), while Hampshire is dual coloured and has texture on the lower side.
Duroc vs Yorkshire	Texture and Colour	Duroc has moderate texture and is largely dark coloured (a few dual coloured), while Yorkshire is fully pink and has almost NIL texture.
Ghungroo vs Hampshire	Texture and Colour	Ghungroo pigs are uniformly high texture and black coloured while Hampshire pigs are dual coloured and have texture on the lower side.
Ghungroo vs Yorkshire	Texture and Colour	On both colour and texture front, these two are poles apart(high and low with respect to texture and black and pink with respect to colour).
Hampshire vs Yorkshire	Colour Alone	Texture profiles are similar. Hence, colour alone is preferred and works as Hampshire pigs are dual coloured while Yorkshire is uniformly pink.

1.6.1 Classification Frame 1: Siphoning Trees and Feature Selectivity

On visual grounds, while scrutinizing the cropped muzzle images from different breeds, there arises a choice of going with Texture or Colour or a combination of Texture and Colour descriptors for separating the breeds. All the possible breed pairs along with the corresponding descriptor by which they are best separated are listed in Table 1.2 and shown graphically in Fig. 1.4. The breed pairs for which the reliable feature type is combination of Colour and

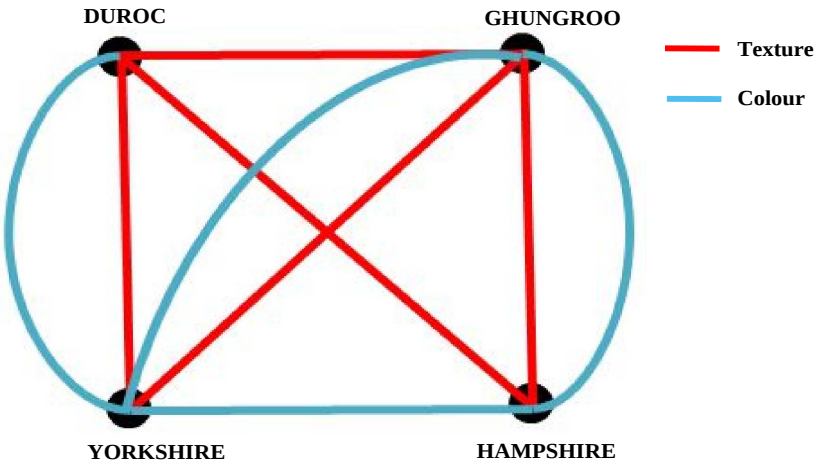


Figure 1.4: Different breed pairs and the feature sets which best separates the pair shown graphically.

Texture descriptors are the ones which are farther apart in feature space as compared to the breed pairs for which the reliable feature type is either Colour alone or Texture alone. As can be observed from this table that all the breed pairs are not separated in the best manner by the same type of descriptor. This provides the motivation to go for a multi-stage classification scheme in a hierarchical fashion. The breed occupying the highest position in the hierarchy is separated out first from the rest. The separation is carried out using that descriptor which best separates the two classes. After one of the breeds has been separated out from the rest, the same process is re-iterated with the remaining breeds and this process continues till all the breeds have been separated. An algorithm has been proposed to determine the order of the hierarchy as well the type of descriptor which is to be used at each stage of classification. This proposed multi-stage classification scheme where the breeds are siphoned out one after another is completely different from a multi-class classification scheme which involves classifying a test muzzle image as belonging to one of the four breeds in a single stage and over a fixed decision space. In the proposed multi-stage classification scheme, at each stage of classification a binary classification is performed using a binary classifier like SVM. Thus, the classification performance at each stage is governed by the virtues and limitations of the SVM.

1.6.2 Classification Frame 2: Data Centric Models based on Spanning Trees

There exists substantial variability in the colour and texture profile of certain breeds, mainly Duroc and Hampshire. This intra-breed variability leads to scattering of the feature vectors constructed from the colour and texture descriptors for a single class or breed. Apart from this inter-breed similarity also exists: for example between Duroc/Ghungroo and Duroc/Hampshire, which leads to overlap between the clusters formed from feature vectors of two different breeds/classes. This intra-breed variability and inter-breed similarity affects the performance of any classification algorithm. For the multi-stage classification scheme discussed in the previous subsection, the performance of the SVM classifiers at each stage of classification is severely affected by these factors. Hence, an attempt is made to develop a classification frame whose performance is less affected in the presence of intra-breed variability and inter-breed similarity and which is not data hungry, i.e. the classification algorithm can provide good performance even with few good quality representative samples.

The proposed classification scheme is a generative model, i.e. it tries to learn robust representation for each of the training classes. For this purpose, representations based on spanning trees are learnt for each class. The proposed classification scheme is a single-stage classifier which uses a common decision space formed from the combination of colour and texture features. Spanning trees discussed in detail in Chapter-6, are useful on many fronts:

- **Emphasis and De-emphasis**: It is important to define what part of the training data is important and is representative of the breed-core and what part is noisy and can be discarded as the samples are of a poor quality. Spanning trees can help achieve this objective.

- **Conflict resolution**: Given the training samples of a particular class, the class specific data-correlation profiles and similarities can be obtained from the Minimum Spanning Tree(MinST) representation. When one attempts to define the spread of class-specific training data, more connected to bounds, Maximum Spanning Tree(MaxST) representa-

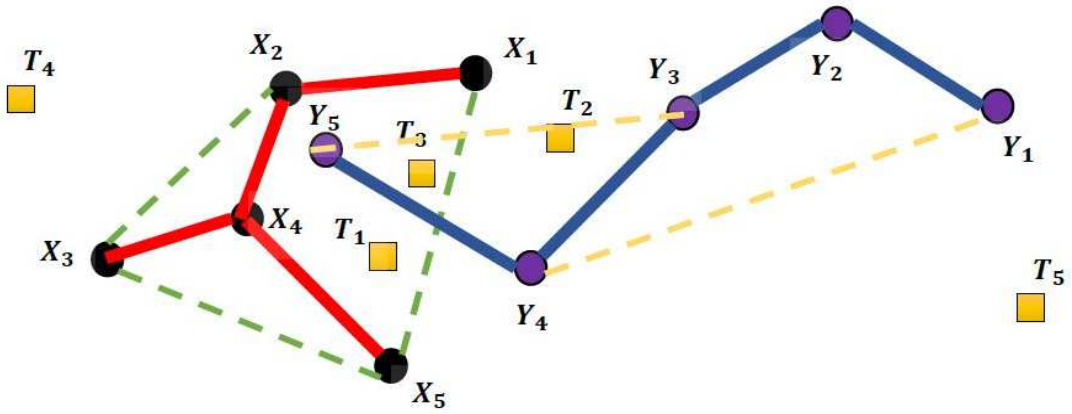


Figure 1.5: MinST model and five test cases.

tion becomes useful. Both these types of trees tend to stitch data-points in their own way so that one can address one of the following questions:

- If a particular test feature vector is embedded inside a training cluster, how deeply embedded is it? How relevant is it in the context of a multi-class problem wherein a given test-point is in the interior overlapping region of multiple clusters from different classes?
- If a test feature vector is in the interior of one of the breed-clusters but is exterior to another, the problem becomes simpler but must be resolvable at a data-centric level, particularly when limited training samples are available for each class. How does a spanning tree frame detect this change in modality?
- If the test feature vector is exterior to all the breed-clusters, then this case becomes simple but the question becomes tricky in the context of the shape and orientation of the cluster. How far is the test point really from each of these clusters?

An exemplar figure comprising of several cases of test feature vector locations and two training clusters represented as MinST-models and MaxST-models are shown below in Fig. 1.5 and

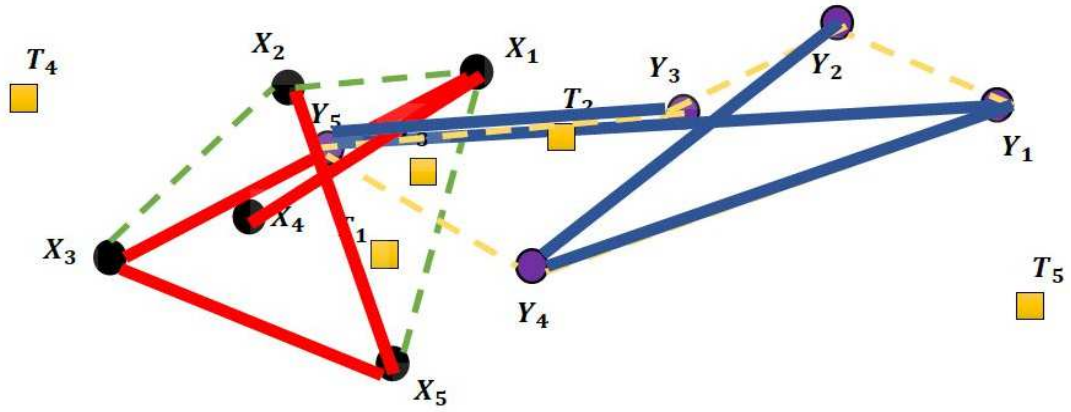


Figure 1.6: MaxST model and five test cases.

Fig. 1.6. Here $T_i, i \in \{1, 2, 3, 4, 5\}$ denote the test feature vector variations, $X_i, i \in \{1, 2, 3, 4, 5\}$ are the training nodes from Class 1 and $Y_i, i \in \{1, 2, 3, 4, 5\}$ are the training nodes from Class 2. To generate a class specific outlier score, the change observed in the MinST and MaxST primary graph(linking the training nodes alone) is quantified after inducting the test feature vector. This in a crude way reveals whether the test feature vector is deeply embedded in one of the class clusters or is outside one or both of them. The outlier score assigned to a test feature vector depends on the amount of change induced by it to the Spanning Tree structure. It has been observed that the farther a test feature vector moves outside the cluster, the higher is the change induced by it to the Maximum Spanning Tree structure formed from the training cluster. On the contrary, if the test feature vector lies inside the training cluster, then the change induced in the Minimum Spanning Tree structure increases if the test feature vector lies closer to any feature vector within the cluster and it is surrounded by more number of neighbours. Thus cases wherein the test feature vector lies in the exterior of all the classes or it lies in the exterior of all classes except one can be resolved using MaxST-models; whereas cases wherein the test feature vector lies in the overlapping region of two or more classes can be resolved using MinST-models.

1.7 Research Challenges, Primary Thesis Contribution and Chapter Outline

The earlier sections bring out the variability with each breed class. In other words, a typical muzzle is not just a homogeneous disk or dual coloured disk. It harbours a variety of structural artefacts coming from the cilia/hair follicles and sweat pores, randomly positioned for each pig within the same breed. Coupled with a natural scene illumination variation, these visible artefacts tend to interfere with the colour profiling process. Thus, since colour profiling alone will not work, it becomes important to use the texture domain as well and indirectly trap parameters such as "density" of artefacts present on the muzzle surface and to some extent their spatial distribution all around the dial. This feature identification problem is exacerbated by lack of availability of sufficient training samples. This makes both the model building and test-point classification processes highly challenging and unique to this application frame. A summary of the research challenges are specified below:

1.7.1 Challenges involved in the breed identification process

The main challenges associated with this breed identification problem are as follows:

- The size of the dataset available for training was highly limited (assuming training and testing is done based on a 50%-50% split and in cross-validation mode to emulate on-field natural variability). This meant that pigs which were used for building the breed-model were completely different from the ones used for testing.

- The natural intra-class variability was found to be very high particularly for dual-coloured pigs such as Hampshire and Duroc. Furthermore, different pigs of the same breed-type exhibited variability with perspective, scale and contour-skew. To add to that there was a variability in the local illumination profile for different pigs as they were snapped in different sheds.

- Each muzzle snapshot (albeit cropped) exhibited some variability in background (in most cases the facial profile of the pig coupled with the glove of the individual holding the

pig's snout formed the background). Separating the muzzle contour from the face of the same pig in the backdrop, was a challenge for some of the breeds such as Ghungroo and Hampshire.

- Numerically 5-7 pigs per breed with 10-variations per pig were available for building a breed-model. This placed severe limitations on the effectiveness of classification models towards multi-class breed-model building.

When the breed classification problem is viewed in relation to the sparse, limited and diverse nature of data available for model building, it is clear that there is a need for a solution which does not over-generalize the construction of the breed-model. That is because, this would imply that breed-types which have similarities in some domain, would be very difficult to resolve, when a test sample falls in the region of intersection of two or more clusters from different breeds in feature space. There are two ways this problem has been handled in this thesis: (i) finding a way to isolate only those features, which impart some form of structure to the breed clusters; and (ii) stitching together the samples within a particular cluster based on some form of association rule with the help of a spanning tree, which makes the construction dataset-adaptive, yet unifies parts of the whole in a certain way.

1.7.2 Contributions of this thesis

In line with the above challenges, the contributions of this thesis are multi-fold and covers the following fronts:

- A customized segmentation algorithm has been designed in which an attempt is made to fit a circular ball shaped segmentation mask with the largest possible radius so that no part of the background region is picked up from the muzzle image for feature extraction. For this purpose suitable filter kernels have been designed and deployed based on differential Gabor variant [23] operator to emphasize the muzzle contour while suppressing the texture details in the muzzle interior. This composite filter is convolution of three identical kernels oriented at $0°, 120°$ and $240°$, which we term as the 'STAR' operator. Using this 'STAR'

operator internal cleaning is done, by considering the centre of the muzzle image as the centre of the circular mask, wherein the median radius is determined from the response of the muzzle image to the 'STAR' operator. This ball fitting segmentation algorithm is muzzle content adaptive.

- Judicious selection of discriminatory statistical measures have been made to segregate breeds so that difficult breeds such as Duroc and Hampshire can be differentiated from the rest. This involved designing suitable texture operators such as GSM [8] and Morphological Tophat operator [1] to enhance textural details in different breeds. On the colour front, statistical parameters such as Sarle's distribution bi-modality index [22], were deployed to detect breeds with dual coloured muzzle images such as Hampshire.

- On the classification front, there are two contributions. The first contribution is centered around the design of a hierarchical classification scheme involving multiple classification stages. The hierarchical classification scheme resembles a tree structure where at each stage of classification hierarchy a particular breed was separated out from the remaining ones in a decision space in which it is optimally separated from the others. The order in which the breeds are to be separated out as well as the decision space at each stage of the hierarchy is decided by our algorithm, which uses a feature specific inter-breed distance table. Over the existing Phylogenetic tree construction using AGNES [24] [25], the proposed siphoning tree was found to provide a compact cluster representation at each decision node, ensuring a much more efficient use of the linear Support Vector Machine (SVM) module.

- The hierarchical classification scheme mentioned above tries to search suitable decision space for each breed where the overlap between classes can be minimized, thus allowing us to achieve the best possible classification accuracy with standard classifiers like SVM. In contrast the second contribution on the classification front tries to improve the classification accuracies further by developing generative models corresponding to each class/breed based on Spanning Trees. The classification model uses an outlier detection framework

based on Maximum Spanning Tree representation for each of the classes/breeds. A novel testing procedure has been proposed to assign class label to a test feature vector corresponding to the class with the lowest outlier score. However, when the test feature vector falls within a cluster of feature vectors from a training class, the Maximum Spanning Tree representation model generated identical outlier scores irrespective of the position of the test feature vector within the cluster; since it could not account for the positional arrangement of the test feature vector with respect to the feature vectors from the training clusters. To take care of this situation, two separate methods were proposed: one based on Minimum Spanning Tree representation and the other involved calculating the outlier score after projecting the feature vectors to $\mathbb{R}^2$ plane using Random Projection Matrices. Although both the methods lead to nearly the same accuracy for the four breeds; but the method based on Minimum Spanning Tree representation is computationally much cheaper as compared to the other. The highlighting feature of our proposed outlier detection framework is that while computing the outlier scores corresponding to a particular training class, the spanning tree model considers the proximity of the test feature vector to the training cluster(of target class as a whole) as well as the positional arrangement of test feature vector with respect to the feature vectors within the training cluster. This has led to the performance improvement of our outlier detection framework as compared to other state-of-the-art methods in literature.

1.7.3 Thesis outline

The outline of the thesis is as follows:

- Chapter 2 discusses a customized adaptive ball-fitting segmentation technique to isolate the muzzle region from the background. Since it is extremely difficult to extract the muzzle contour precisely even with state-of-the-art segmentation techniques in literature, hence instead of going for an over-estimate of the muzzle contour, an attempt is made to fit a circular ball shaped segmentation mask inside the muzzle contour with the largest possible radius, so that no part of the background region is picked up. The selection

and design of a suitable pre-filtering operator followed by the design of an algorithm to calculate the radius of the circular segmentation mask is discussed. The effect of background leakage on the texture statistic is also discussed in this chapter.

- Chapter 3 discusses the colour and texture descriptors followed by the construction of feature vectors which can impart distinctiveness to each of the four breeds. These feature vectors are constructed based on the colour and texture descriptors from the segmented muzzle region only. Two different texture descriptors are described: one based on the GSM and the other based on the Morphological Tophat Transformation operator. Suitable attributes are defined and aggregated from the response of a muzzle image to these two operators to form the texture feature vectors. Similarly the 2D $C_b - C_r$ histogram constructed from the description of the image in the $YCbCr$ domain acts as the colour descriptor. Suitable statistics based on the footprint of the $C_b - C_r$ histogram are defined and aggregated to form the colour feature vectors.

- Chapter 4 discusses a multi-stage classification scheme for breed classification with three feature sets viz. Colour(C), Texture(T) and combination of Colour and texture($C \cup T$). At each stage of classification, a single breed is siphoned out from the rest using a SVM classifier in a suitable feature/decision space as decided by the proposed classification algorithm. First of all, a distance metric between feature vectors from two different classes is defined by treating feature vectors from any class a single entity. With this distance metric, and corresponding to four different breeds, a total of $\binom{4}{2} = 6$ pairwise distances are calculated for each of the three feature types (C), (T) and $(C \cup T)$. These distance metrics are fed as the input to our multi-stage classification algorithm to determine the order of breed-siphoning as well the feature types to be used at each stage of siphoning. The performance comparison of our proposed multi-stage classifier which generates a tree-like classification hierarchy is compared with other tree based classification methods in literature.

- Chapter 5 discusses a multi-class classification scheme based on an outlier detection frame-

work using Maximum Spanning Trees and Random Projection Matrices which can provide high classification accuracies even in the presence of intra-breed variability and inter-breed similarity. These two problems severely affected the classification performance of SVM in the multi-stage classification scheme discussed in Chapter 4. To tackle the problem of intra-class variation, a formulation based on Minimum Spanning Tree(MinST) is presented which splits any training class into multiple clusters and enables learning Maximum Spanning Tree(MaxST) for each such cluster within a class. The MaxST representation of a training class along with the method of testing is discussed in detail. To account for the scenario, when the test feature vector lies inside the cluster of training feature vectors, a method which involves calculating the outlier score after projecting the feature vectors to $\mathbb{R}^2$ plane using Random Projection Matrices is discussed. Outlier scores corresponding to each training class is generated for a test feature vector. The method for making inference about breed label for the test feature vector from the outlier score corresponding to each class followed by the performance comparison of our proposed method with standard multi-class classifiers like SVM, Decision Trees, Nearest Neighbour algorithms is also discussed.

- Chapter 6 discusses an improvement of the outlier detection mechanism discussed in Chapter 5. The method which involves calculating the outlier score after projecting the feature vectors to $\mathbb{R}^2$ plane using Random Projection Matrices is found to be computationally intensive. Hence, this method is replaced by another method which takes advantage of the MinST representation of the clusters within a training class. The MinST representation is found to be sensitive to the positional arrangement of the test feature vector with respect to the training feature vectors of the training cluster within which it lies. The performance comparison of this modified outlier detection framework with respect to a target class which uses MinST representation in conjunction with the MaxST representation is compared with state-of-the-art classification methods in literature. The multi-class classification strategy developed using this modified outlier detection framework is discussed and its performance comparison is done with respect to the method

discussed in Chapter 5.

- Chapter 7 concludes the thesis with a summary of the research work carried out and an outline of the future research.

2

Adaptive Ball Fitting Segmentation Algorithm

Contents

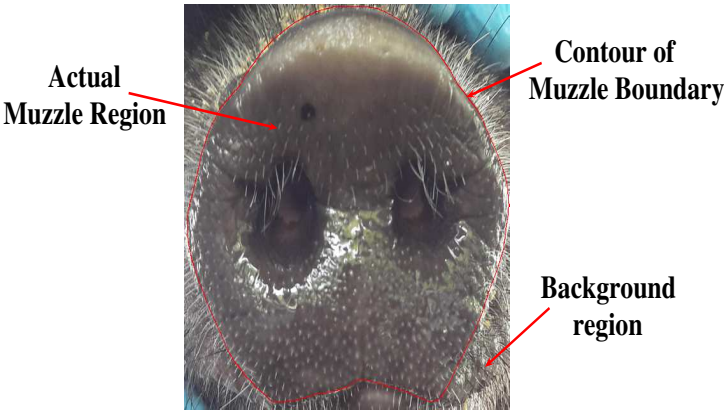


Figure 2.1: Cropped muzzle image with the actual muzzle region shown separated from the background by the contour of the muzzle boundary.

As discussed in Chapter 1, one of the reasons why the muzzle region was chosen as a biometric for breed analysis, was because it was a part of a sensory interface. Behavioural traits over generations tend to influence the shape, colour, ruggedness, roughness and to a large extent the texture profile of the muzzle of a typical pig from a particular breed. From a visual standpoint, one could use any one or a combination of the following descriptors to profile the breed-type: Shape, Colour or Texture. The muzzle contours of Hampshire and Yorkshire are mostly shaped in the form of a cardioid [26], while the Ghungroo breed has a largely circular shaped muzzle. All these three breeds have notches at the bottom of the muzzle contour. Apart from these three breeds, Duroc has the most distinct shape, which is other than that of a cardioid or circular. Most Duroc pigs exhibit two prominent inward notches at around 60° and 120°, which are not present in the other three breeds. Despite this distinctiveness, due to background interference and perspective variations(rotation and skew changes), precise segmentation of the muzzle contour is very difficult.

While muzzle shape analysis can be dismissed as a breed-identifier, it is essential to extract only that part of the muzzle interior from which texture and colour features and statistics can be computed(either locally or globally). The question that arises is how does one specify the region of interest. It should ideally fall in line with the muzzle contour or periphery and the entire interior should be used for analysis. However, as will be shown in the initial part of this

chapter, this is not a good idea. Going ahead with an attempt of over-precise segmentation process can result in involving a significant portion of the background. Since the texture density in the background region is usually on the higher side and the background colour profile is unpredictable, its inclusion towards final feature and statistic computation is likely to increase the standard deviation of the of the colour and texture statistics; thus increasing the intra-class variability.

The muzzle database available with us has cropped muzzle images as shown in Fig. 2.1, which comprises approximately 25% as background area. While extracting breed specific features from the muzzle surface, this background consisting of some parts of the facial region of the pig along with the surrounding region presents some interference, thus corrupting the feature vectors. Thus it becomes important to isolate the muzzle area from the background prior to feature extraction. The graphical flow of the entire breed identification process can be explained with the help of Fig. 2.2.

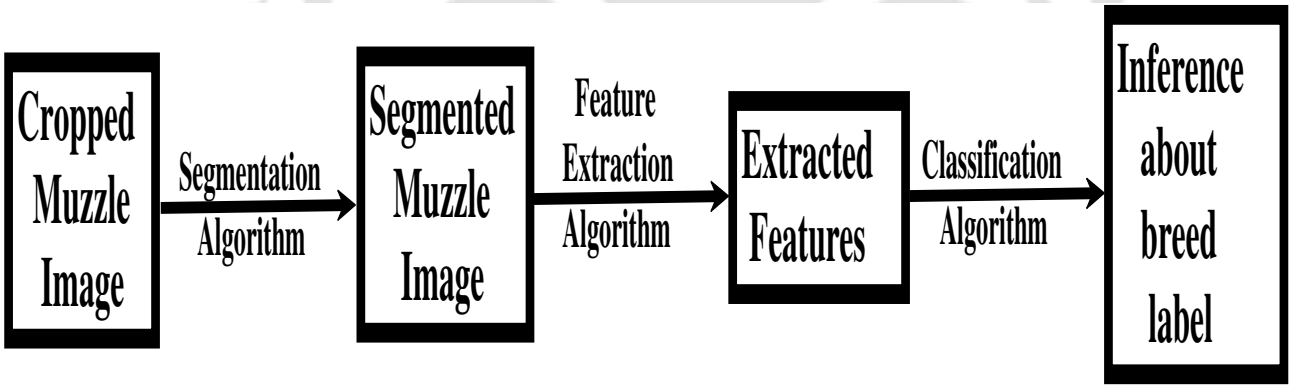


Figure 2.2: Overall block diagram of the classification pipeline.

Thus, given a cropped muzzle image, the first step, is to extract or estimate the muzzle contour from a noisy background which is both rich in colour as well as texture. Feature extraction can be done once this region of interest is detected and is confined to the interior of the muzzle. The role of a segmentation process is to ideally extract out the muzzle region from the background for further processing and feature extraction. Active contours using level sets are one of the most extensively used methods in image segmentation [27, 28]. This is due to their inherent ability to adapt to complex contours after a certain number of iterations, thus

defining the boundary between regions in an image. The level set methods in literature are broadly categorized as edge based or region based.

In edge based level set methods, minimizing the energy functional for curve evolution is equivalent to locking the contour onto edges in an image. These methods are suitable when the object to be segmented is separated from the background by sharp edges. The distance regularization term in [29] plays a crucial role in locking the contour to complex boundaries with high edge strength. While in region based level set methods, the minimization of energy functional is equivalent to segmenting the image into homogeneous regions. The measure of homogeneity could be based on gray level intensity or colour or some texture-profile. In an early work by Chan et al. [30] involving region based level sets, a degradation in performance was observed when there was an intensity in-homogeneity either in the foreground or the background. Wang et al. [31], [32] later covered this issue, so long as this intensity in-homegeneity was found to be within certain bounds. Cai et al. [33] on the other hand used colour information guided by visual saliency for segmentation. Zhi et al. [34] used a combination of edge and region based methods.

However, these methods were not suitable for pig muzzle segmentation, due to the following practical on-field constraints:

- **Colour-profiling issues:** When the snapshot of the muzzle is taken from the front, the region immediately around the muzzle contour is usually either the face of the pig and/or partially a glove (belonging to the individual holding the pig to stabilize the head movement). Since the location of the glove around the muzzle(if present) is likely to vary and furthermore, in some cases this may not even be picked up in the photograph, there will be an inconsistency both in the background colour profile as well as the texture profile around the muzzle periphery. This inconsistency in background interference, if picked up during the segmentation procedure will severely impair the stability of the features, both on the colour and the texture front. In the initial part of this chapter, we illustrate this sensitivity on the texture front alone, when a small portion of the background is erroneously recognized as the segmented region.

- **Texture-profiling issues:** From a texture viewpoint, the presence of lines, edges, corners, curves in the form of pores and hair spikes over the muzzle surface, impart a certain texture profile to the interior of the muzzle. However, the background around the muzzle periphery, if included will also contain some artefacts and in fact in most cases will also be rich in texture(although may have a different composition as compared to the muzzle interior). Since the background scene during muzzle photography will vary from pig to pig within a breed, there will be a large intra-class variability with respect to texture as well. Given a broad definition such as this, it becomes difficult to segregate the background containing the face of the pig which also has a rich texture due to furry and spiky hair present over this facial region.

Since standard segmentation algorithms fail, feature mixing is virtually unavoidable if one attempts an over-precise detection of the muzzle contour and its interior. Hence, instead of going for a precise contour extraction, an attempt is made to fit in a ball inside the muzzle (with the largest possible radius) as shown in Fig. 2.3, so that no part of the background is picked up. The initial goal is therefore to pre-filter the muzzle portion so that the interior gets cleaned up as much as possible and becomes largely homogeneous in appearance. While this interior is being cleaned up, the signal enhancement algorithm also enhances the contour linings without over emphasizing other parts. The segmentation algorithm must therefore achieve the following dual objective:

Selection and tuning of a suitable pre-filtering operator in a way that the muzzle contour is brought out clearly (or enhanced), while at the same time suppressing the details in the interior, making it appear partially homogeneous with respect to texture.

Once this selective enhancement is done, an attempt is made to fit in a circular mask that gets nicely inscribed in the interior of the muzzle. The circular mask must be large enough so that it can trap sufficient details both with respect to colour and texture during the feature extraction procedure. If the radius of this mask is too small (i.e. a conservative strategy to ensure the mask covers only the interior), the fraction of the muzzle area available for analysis will also be very less and this will eventually reduce the efficacy of the breed characterization

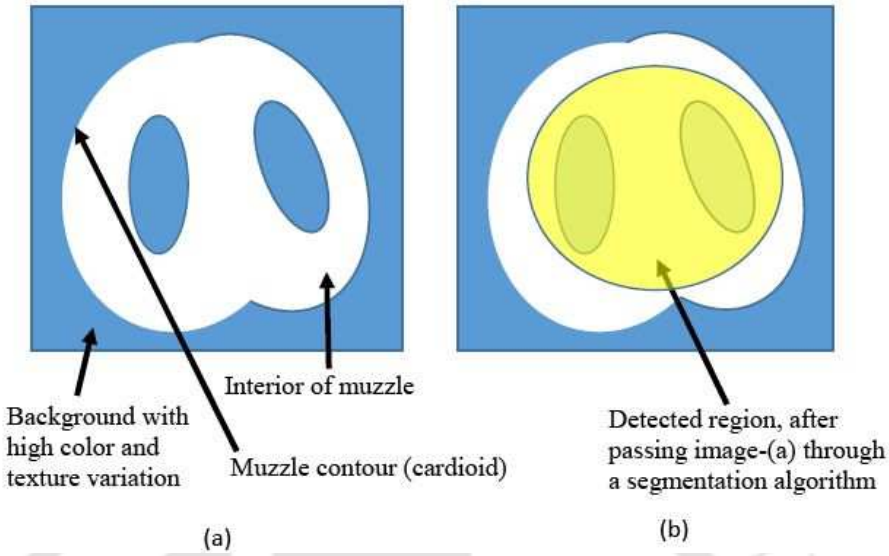


Figure 2.3: Segmentation procedure: (a) Cropped muzzle image; (b) Detected region containing only muzzle interior.

process. On the other hand if the radius is too large this may pick up unnecessary noise associated with the background and boundary variations. Hence, the search for the optimal radius (keeping this trade-off in mind) is done over a certain band R_{MIN} and R_{MAX}. It may be noted that a part of the data in the interior of the muzzle is also lost (rather of less use), owing to the presence of two large nostrils in the muzzle interior. Hence, the minimum radius R_{MIN} must be at least large enough to go beyond the nostrils. Fig. 2.4 shows the muzzle region segmented out from the background using a variety of methods such as an edge based active contour model DRLSE [29]; a region based active contour model [30]; and a combination of region and edge based active contour model SDREL [34] along with our proposed method;. Active contour models which try to maintain region homogeneity by minimizing the energy functional fail miserably for cases where the muzzle region contains the pinkish white patch embedded within the grey region. These methods also fail to differentiate the facial region from the muzzle, when their colour is almost the same. The edge based methods suffer from the problem of boundary leakage, where the edges are very weak. In the following the pre-filtering operator used to enhance the muzzle contour followed by the mask generating algorithm are discussed.

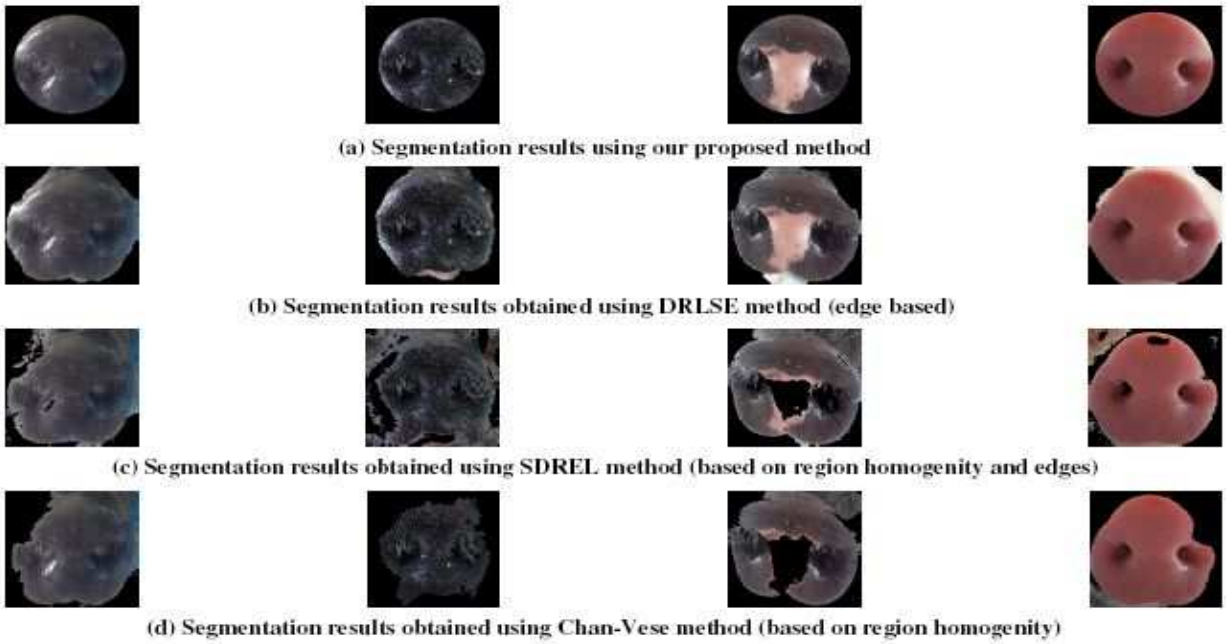
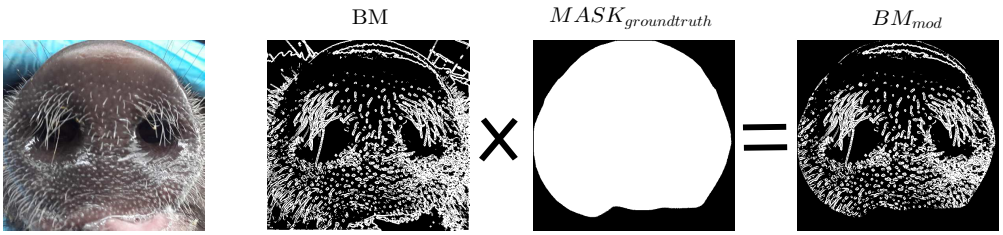


Figure 2.4: Segmented region obtained using (a)Our proposed algorithm (b)edge based DRLSE model (c)Combined region and edge based SDREL model and (d)Region based Chan-Vese model

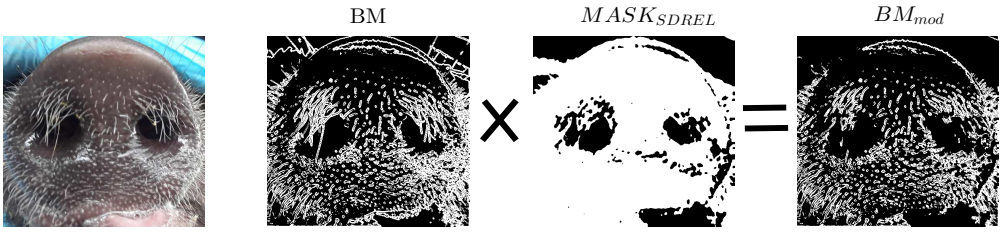
2.1 Muzzle Contour Sensitivity Analysis

The effect of background interference on the extracted features can be understood by measuring the sensitivity of the amount of textural details present in a muzzle image. Consider the binary maps generated in Fig. 1.3. Let BM denote such a binary map. The foreground(or white pixels) in such a binary map highlight the location of textural details in the image. Let $MASK$ denote the binary segmentation mask which is used to isolate the muzzle region from the background. It is the textural details in BM falling under the segmentation mask $MASK$ which is to be used for making inference about breed label. Let BM_{mod} denote the binary map obtained after modulation of BM by $MASK$. The modulation of BM by the segmentation mask $MASK$ to obtain BM_{mod} is shown in Fig. 2.5.

Let n_s denote the number of foreground pixels in BM_{mod} whose size is $N_1 \times N_2$. A statistic S is defined which denotes the total amount of textural details in the muzzle region defined by the segmentation mask $MASK$ normalized with respect to the size of the image. Then



(a) Segmentation with actual mask. $S = 0.2094$



(b) Segmentation with SDREL mask. $S = 0.3081$

Figure 2.5: Figure showing an instance of the change in statistic S, when improper mask is applied.

$$S = \frac{n_s}{N_1 \times N_2} \tag{2.1}$$

$$n_s = \sum_{x=1}^{N_1} \sum_{y=1}^{N_2} BM_{mod}(x, y)$$

$$BM_{mod}(x, y) = BM(x, y) \times MASK(x, y) \forall x, y$$

In order to measure the sensitivity of S as a function of error in the mask generated by any segmentation algorithm with respect to the actual ground truth segmentation mask, a set of 20 muzzle images, with 5 muzzle images per breed as shown in Fig. 2.6 are chosen. The aim is to measure the sensitivity of S breed-wise, i.e. we want to study the effect of mask error as a function of the different breeds. Let $M_{BR}(i)$ denote the statistic S obtained by the application of manually extracted ground truth segmentation mask on the $i^{th}, i \in \{1, 2, .., 5\}$ muzzle image from breed $BR, BR \in \{DUROC, GHUNGROO, HAMPSHIRE, YORKSHIRE\}$. The manually extracted ground truth segmentation masks for the muzzle images in Fig. 2.6 are shown in Fig. 2.7.

Let $S_{BR}^{MED}\{ORIG\}$ denote the representative S value from breed BR when the statistics are calculated using the ground truth segmentation masks. Then $S_{BR}^{MED}\{ORIG\}$ is defined as

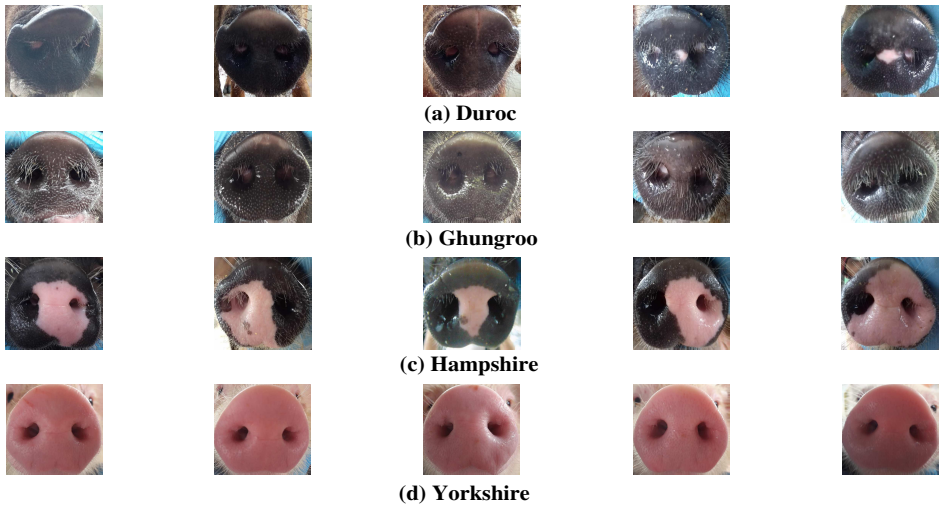


Figure 2.6: Figure showing a total of twenty muzzle images, with five muzzle images per breed.

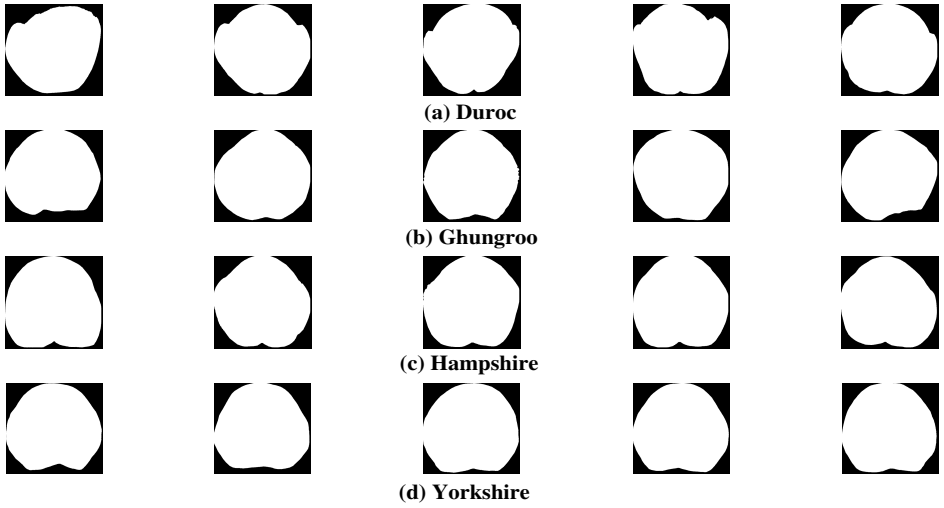
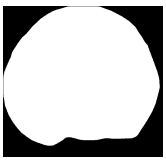


Figure 2.7: Manually extracted segmentation masks for the muzzle images in Fig. 2.6.

$$S_{BR}^{MED}\{ORIG\} = MEDIAN\{M_{BR}(i)\} \tag{2.2}$$

The effect of error in the segmentation mask generated by the different active contour models as mentioned in Fig. 2.4 on the sensitivity of the statistic S is studied. Let $M_{BR}^{AC}(i)$ denotes the statistic S obtained by the application of the mask generated by the active contour method $AC, AC \in \{DRLSE, Chan-Vese, SDREL\}$ on the muzzle image. If $S_{BR}^{MED}\{AC\}$ denotes the



(a) Groud Truth Segmentation Mask

(b) Noisy Versions of Groud Truth Segmentation Mask

Figure 2.8: Ground truth segmentation masks and its noisy versions.

representative S value for breed BR, then

$$S_{BR}^{MED}\{AC\} = MEDIAN\{M_{BR}^{AC}(i)\} \tag{2.3}$$

The sensitivity with respect to the active contour method AC is then defined as

$$SENS_{BR}^{AC} = |log[\frac{S_{BR}^{MED}\{AC\}}{S_{BR}^{MED}\{ORIG\}}]| \tag{2.4}$$

The breed specific sensitivity of the statistic S is also studied by corrupting the ground truth segmentation masks in Fig. 2.7 with noise. Corresponding to every segmentation mask in Fig. 2.7, a set of noisy versions are generated as shown in Fig. 2.8. Let the statistic S obtained by the application of these noisy masks be denoted by $M_{BR}^{NM}(i,j)$, where $i \in \{1,2...5\}$ as earlier and $j \in \{1,2...30\}$. Thus corresponding to breed BR, if the representative statistic of S is denoted and defined as

$$S_{BR}^{MED}\{NM\} = MEDIAN\{M_{BR}^{NM}(i,j)\} \tag{2.5}$$

then the sensitivity with respect to noisy masks is defined as

$$SENS_{BR}^{NM} = |log[\frac{S_{BR}^{MED}\{NM\}}{S_{BR}^{MED}\{ORIG\}}]| \tag{2.6}$$

Experiments have also been performed to observe the change in the statistic S when the ground truth segmentation masks in Fig. 2.7 are eroded by small amounts. Thus for each of the masks in Fig. 2.7, we generate multiple eroded masks. For any random mask selected from Fig. 2.7, the multiple eroded masks are generated by the following procedure. After the centroid of the original mask is located, the average radius of the mask is computed. Let this average radius be denoted by R_{mean}. Five different disk shaped structuring elements with radius $r_i = i \times 0.01 \times R_mean, i \in \{1, 2, ...5\}$ are then generated and used to erode the original mask, thus generating five different eroded mask. Let the statistic S obtained by the application of these eroded masks be denoted by $M_{BR}^{ER}(i, j)$, where $i, j \in \{1, 2...5\}$. Similar to the earlier cases, corresponding to breed BR, if the representative statistic of S is denoted and defined as

$$S_{BR}^{MED}\{ER\} = MEDIAN\{M_{BR}^{ER}(i, j)\} \tag{2.7}$$

then the sensitivity with respect to noisy masks is defined as

$$SENS_{BR}^{ER} = |log[\frac{S_{BR}^{MED}\{ER\}}{S_{BR}^{MED}\{ORIG\}}]| \tag{2.8}$$

The sensitivity values $SENS^{AC}$, $SENS^{NM}$ and $SENS^{ER}$ are tabulated in Table 2.1. The sensitivity values in Table 2.1 shows that even if the segmentation mask shrinks by a small amount with respect to the ground truth segmentation mask, so that it does not cover any portion of the background region, the statistic S do not change much. Thus, these values provides a strong support in favour of our proposed adaptive ball based segmentation algorithm, where we try to fit a circular segmentation mask with the largest possible radius so that no part of the mask penetrates outside the muzzle contour. Also it can be observed from the table that the sensitivity is quite high for Yorkshire as compared to the other three breeds. This is because of the fact that the binary map contains a lot of response near the muzzle contour, whereas there is virtually no response inside the muzzle contour. Therefore, as the

Table 2.1: Table showing the sensitivity values of statistic S for the segmentation mask generated using different methods as a function of the four breeds viz. DUROC(D), GHUNGROO(G), HAMPSHIRE(H) and YORKSHIRE(Y).

	Duroc	**Ghungroo**	**Hampshire**	**Yorkshire**
$SENS^{ER}$	0.0398	0.0174	0.0146	0.5625
$SENS^{NM}$	0.44	0.1298	0.2343	0.6234
$SENS^{Chan-Vese}$	0.1841	0.3577	0.2619	0.7399
$SENS^{DRLSE}$	0.1671	0.0219	0.1092	0.5208
$SENS^{SDREL}$	0.3661	0.1086	0.0344	0.5741

ground truth segmentation mask erodes even by a small amount it loses a significant amount of textural details coming from the muzzle periphery and thus the S value changes by a large amount. Moreover, these textural details on the periphery for the are not important from the breed classification point of view. It is the internal details of the muzzle that are of paramount importance and hence needs to be retained by the segmentation mask.

2.2 Pre-filtering operator used for contour enhancement

Orientation-specific Gabor filters have been used extensively to detect texture patterns in literature [23, 35, 36]. The motive for modifying and adapting these directional Gabor filters, with differentiation along one direction and smoothing along the orthogonal direction, was to enhance the step edges, trapping parts of the main contour. The smooth regions are expected to register a lower detection-score as compared to the edges. In order to enhance the muzzle contour as compared to the smooth interior a quantized Derivative of a Gaussian (DGau) function is presented in its discretized form as:

$$D(x) = -xe^{-x^2/(2\sigma_f^2)} \tag{2.9}$$

Here, σ_f is the standard deviation associated with this Gaussian and $x \in \{-L, .., 0, ..., L\}$, with $L = \lfloor 3\sigma_f \rfloor$ (with $\lfloor \rfloor$ representing the floor function). The base kernel is created by rotating the discretized version in the X-Y plane. The kernel has two degrees of freedom: (i) Standard deviation σ_f associated with the differentiation process along a certain line and (ii) thickness

t associated with a smoothing along a direction orthogonal to the orientation of the DGau function. If $\delta_{1D}(m)$ represents the 1D Kronecker delta function and $\delta_{2D}(m,n)$ represents the 2D Kronecker delta function, then they are defined as follows:

$$\delta_{1D}(m) = 1 \text{ if } m = 0 \text{ and '0' if } m \neq 0 \tag{2.10}$$

$$\delta_{2D}(m,n) = 1 \text{ if } (m = 0, n = 0) \text{ and '0' if } (m \neq 0 or n \neq 0) \tag{2.11}$$

Both $(m,n) \in Z$, the set of integers. The 2D linear convolution denoted by $h(x,y)$ between two functions $h_1(x,y)$ and $h_2(x,y)$ is defined as:

$$\begin{aligned} f(x,y) &= h_1(x,y) * h_2(x,y) \\ &= \sum_{u=-\infty}^{\infty} \sum_{v=-\infty}^{\infty} h_1(u,v) h_2(x-u, y-v) \end{aligned} \tag{2.12}$$

The proposed pre-filtering operator has two wings: one a derivative filter and the other a smoothing filter in the orthogonal direction. This operator is obtained by a 2D linear convolution between two functions $h_1(x,y)$ and $h_2(x,y)$, defined as follows:

$$h_1(x,y) = D(x); for y = 0 \tag{2.13}$$

Written compactly as,

$$h_1(x,y) = \sum_{u=-\infty}^{\infty} D(u) \delta_{2D}(x-u, y) \tag{2.14}$$

The smoothing wing is given by:

$$h_2(x,y) = \sum_{u=-\frac{(t-1)}{2}}^{\frac{(t-1)}{2}} \delta_{2D}(x, y-u) \tag{2.15}$$

$$h_{c,0F}(x,y) = h_{BASE}(x,y) = h_1(x,y) \star h_2(x,y) \tag{2.16}$$

This is equivalent to running the first filter h_1 over the entire image and following it up with h_2 (or vice-versa). The first filter which is the DGau filter, eliminates zones where the intensity profile is largely homogeneous and converts step edges to thick lines. The second filter, depending on the thickness parameter t, serves as a brush (whose thickness is decided by t) in extending the impact of the directional gradient over a small neighbourhood.

Let (X, Y) be the new coordinates when the reference system is rotated in the counter-clockwise direction by an angle θ and let (x, y) the coordinates with respect to the original reference system.

$$\begin{aligned} x &= X\cos(\theta) - Y\sin(\theta) \\ y &= X\sin(\theta) + Y\cos(\theta) \end{aligned} \tag{2.17}$$

And in this new reference frame,

$$h_{c,120F}(x, y) = h_{BASE}(X = x\cos(\theta) + y\sin(\theta), Y = -x\sin(\theta) + y\cos(\theta)) \tag{2.18}$$

where $\theta = 120°$ in the above equation. A similar pattern is followed for the other function positioned at 240 degrees. The generated plots are shown in Fig. 2.9. There are in fact three primary, directional DGau-type filters in operation, whose results are fused using the square energy linked relation [23]. The final texture profile is,

$$\begin{aligned} I_F(x, y, \theta) &= IM(x, y) \star h_{c[\theta]F}(x, y) \\ T(x, y) &= [I_F(x, y, 0^0)]^2 + [I_F(x, y, 120^0)]^2 + [I_F(x, y, 240^0)]^2 \end{aligned} \tag{2.19}$$

In order to enhance the relevant textural details and eliminate the irrelevant ones which are oriented along different directions, two rotated versions of the operator $h_{c,0F}$ are created. When the orientation of this kernel is changed along the X-Y plane to 120^0 and 240^0 (Fig. 2.9), this definition of homogeneity extends to include elimination of lines at other orientations. Curves

can also be eliminated provided the curvatures at points where there is a change in direction is not significant. Pores are also eliminated.

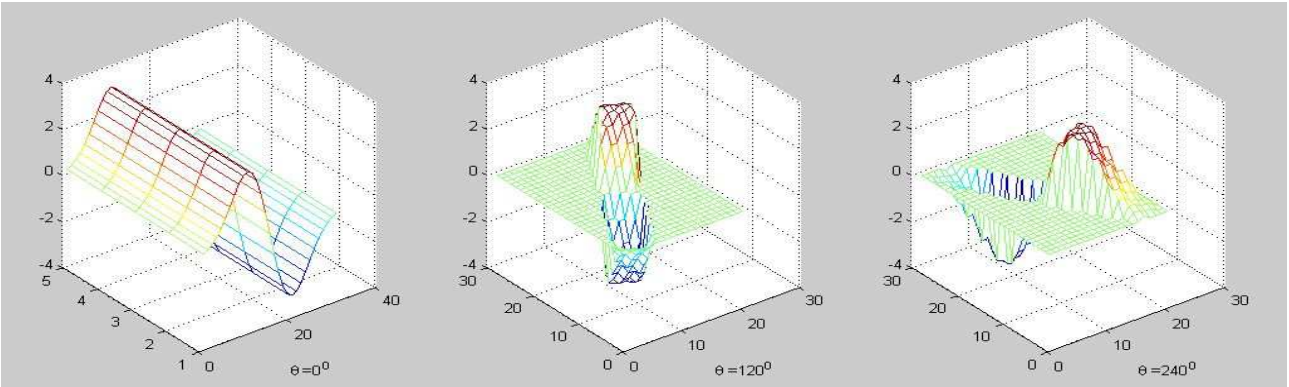


Figure 2.9: Structure of the directional (orientation specific) DGau filters corresponding to three different angles: $0^0, 120^0$ and 240^0, all with respect to the X-axis.

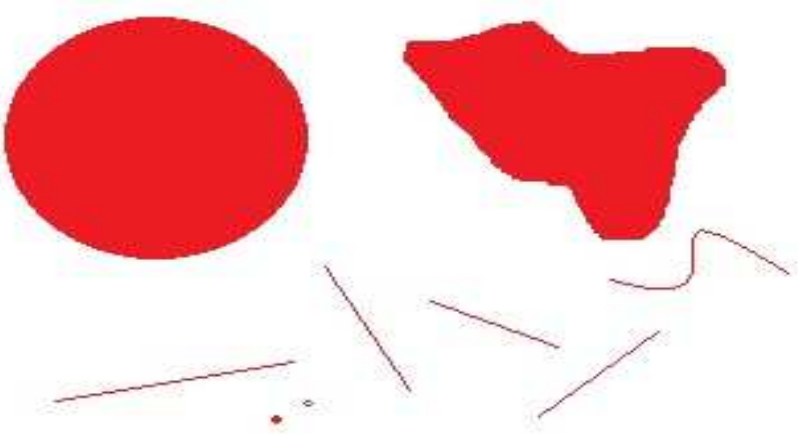


Figure 2.10: Toy texture patterns for testing the effectiveness of the filter chosen for analysis.

An increase in σ_f tends to thicken the contour profiles as the DGau function tends to operate along the normal to the contours. To apply this design, it is recognized that the muzzle pre-filtering procedure has two degrees of freedom:

- **Standard deviation,** σ_f The strength of the orientation specific Gabor-differentiator can be adjusted by increasing or decreasing σ_f. Because of the circular symmetry of triad-arrangement $(0^0, 120^0, 240^0)$, most line artifacts owing to the presence of hair in

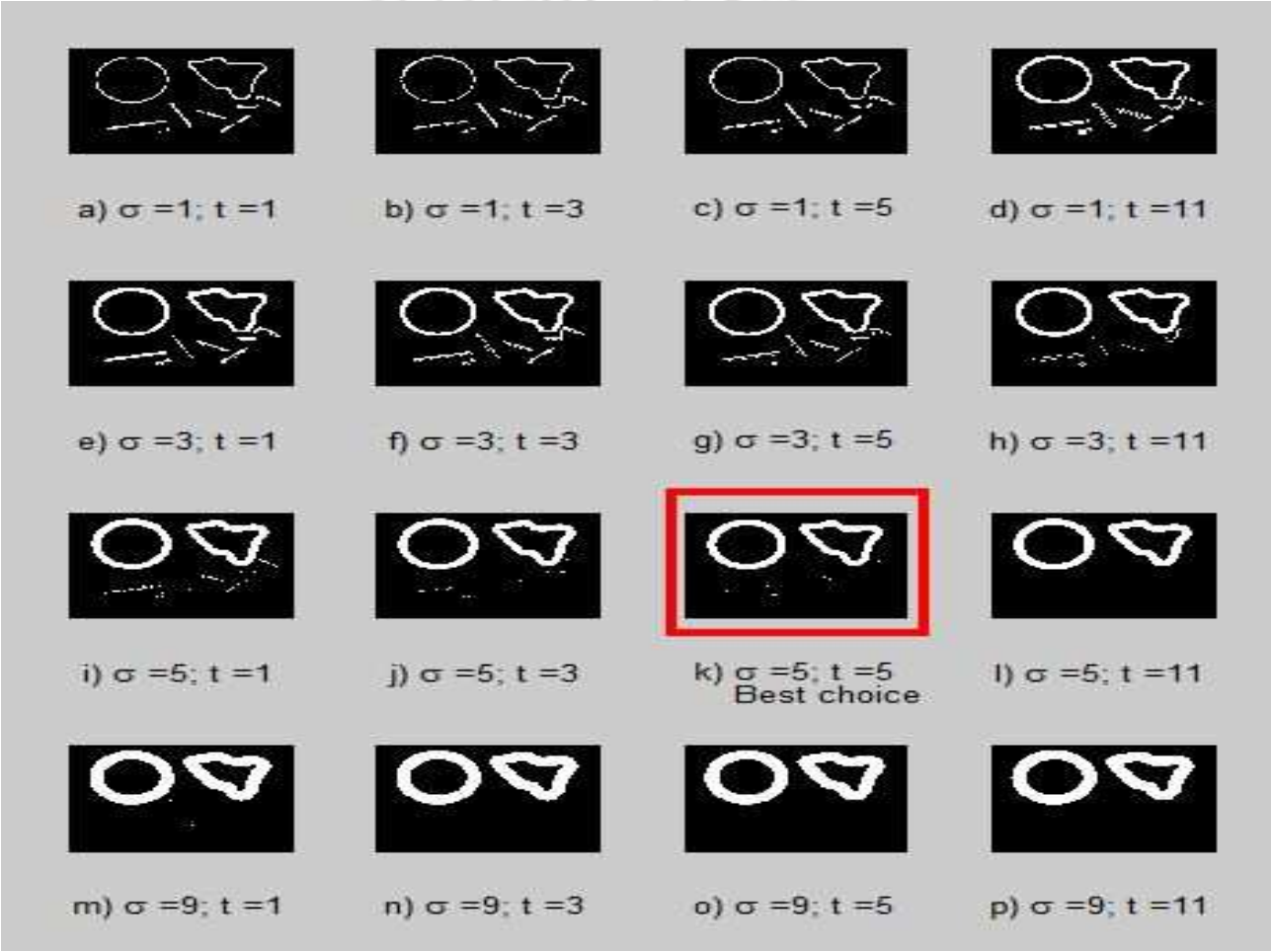


Figure 2.11: Results of DGau operation for different parameter values (σ_f, t) on the toy image of Fig. 2.10. Preferred result is boxed.

the muzzle interior can be eliminated. This triad-arrangement is fixed (inclusion of more spokes in this wheel will only add redundancy to this form of profiling). The interior of the muzzle has many sweat pores (in some cases covered by stunted hair), which emerge as spikes. These spikes can be suppressed is the value of σ_f is sufficiently large (illustrated in Fig. 2.13 with respect to the muzzle of the Ghungroo pig in Fig. 2.12). When this DGau is aligned with a line artefact, the line can be removed completely. But since the hair on the muzzle may have arbitrary orientations, this line elimination is indirectly facilitated through a projection mechanism involving three DGau along three different directions (indicated by the filters in Fig. 2.9).

- **Thickness,** t The effectiveness of the brush work involved in cleaning up the interior increases with t. This also increases the tolerance of the triad-DGau structure to multiple line thicknesses, sweat pores of different sizes and arbitrary short curves. But too large a thickness t has a tendency to enhance step edges particularly in the case of breeds which have pink-patches in the interior, such as Hampshire and Duroc. Furthermore, this increase in t may lead to an overemphasis of the contours associated with the pig's nostrils (which is unnecessary).

Impact of selection of the filter parameters on a toy-image containing some simple patterns in Fig. 2.10 are shown in Fig. 2.11 for different parameter values: $\sigma_f \in 1, 3, 5, 9$ and $t \in 1, 3, 5, 11$. An increase in σ_f tends to thicken the contour profiles as the DGau function tends to operate along the normal to the contours.

To crystallize the parameters for our breed classification problem, calibration is done with respect the most noisy breed, which happens to be Ghungroo and then test the final set on all. To identify the right choice of parameters, σ_f is varied over the set $\{1, 3, 5, 9\}$ and the thickness is varied from $1, 3, 5, 7$. The original cropped Ghungroo muzzle image is shown in Fig. 2.12 upon which the impact of the Gabor filtering followed by the binarization procedure is witnessed. High texture homogeneity within the Ghungroo-muzzle contour is obtained for $\sigma_f = 4; t \geq 3$, as indicated in the sub-figures inside the red-rectangle in Fig. 2.13.

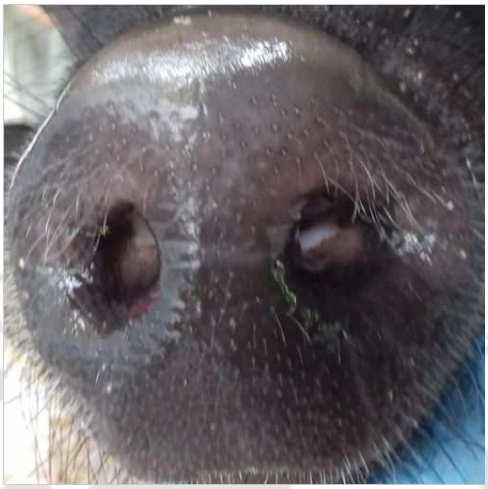


Figure 2.12: Muzzle of Ghungroo pig for DGau pre-filtering and texture quantization.

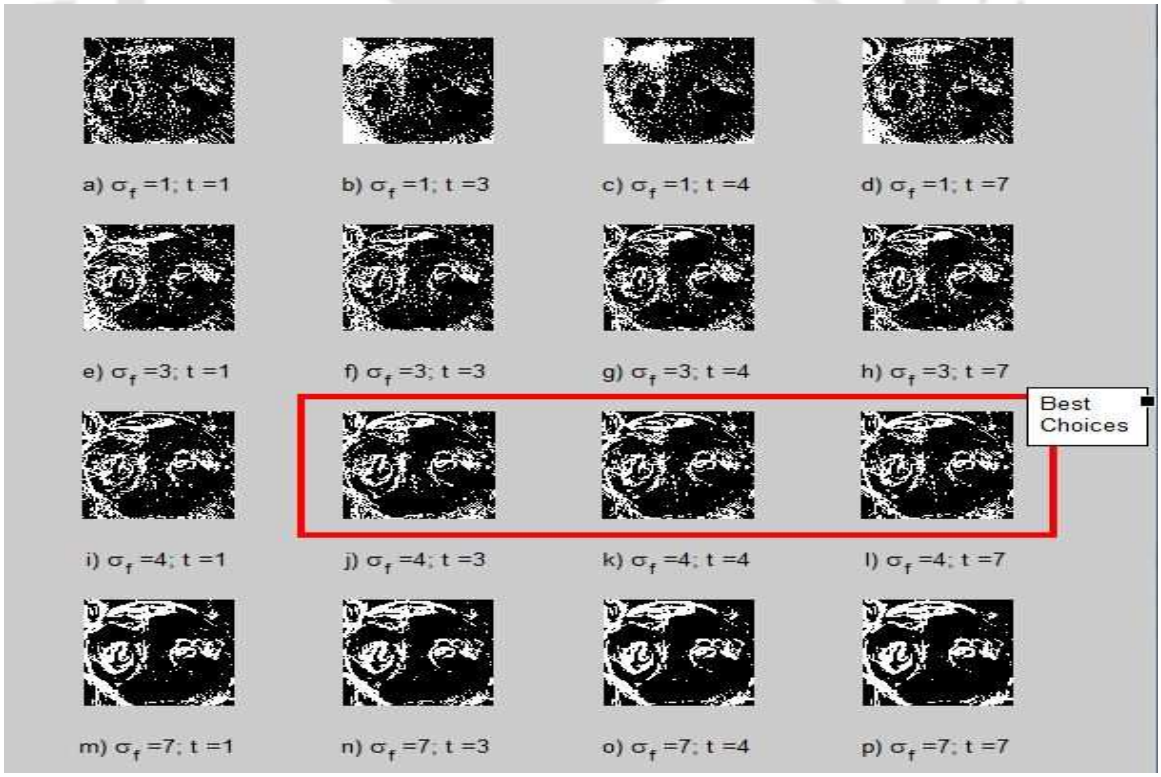


Figure 2.13: Impact of DGau pre-filtering and texture quantization on the muzzle of a Ghungroo pig (Fig. 2.12).

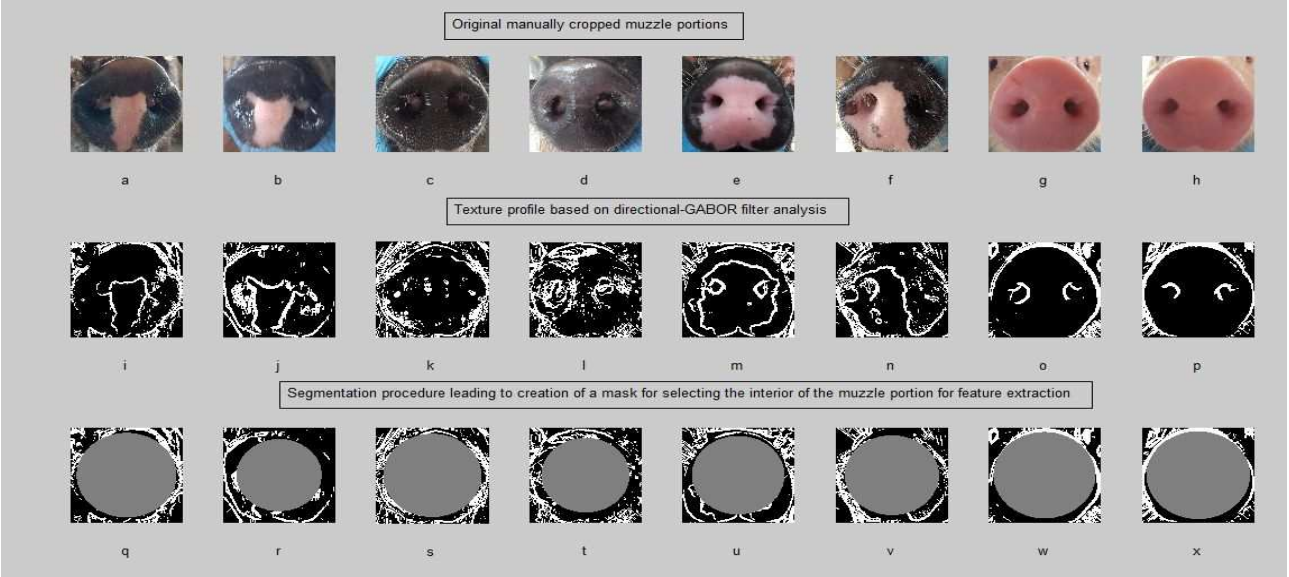


Figure 2.14: Results of adaptive circular mask generation process for pigs from different breeds.

2.3 Generation of the circular mask

Once the response of the muzzle image to the pre-filtering operator has been obtained, the circular mask needs to be generated from this response. The centre of the circular mask coincides with the centre of the cropped muzzle image and the radius of the circular mask is what needs to be determined. The process starts with the computation of the global mean of $T(x, y)$ as

$$\mu_{GL} = \frac{1}{N^2} \sum_{x=1}^{N} \sum_{y=1}^{N} T(x, y) \tag{2.20}$$

and the final quantized binary representation is given by,

$$BIN_{TEX}(x, y) = 1 \text{ IF } T(x, y) > \mu_{GL} \text{and '0' otherwise} \tag{2.21}$$

for $x, y \in 1, 2, ..N$ $(N = 512)$. Once the binary segmentation map $BIN_{TEX}(\sigma_f, t)$ is created for the best choice of parameters $\sigma_f = 4$ and $t = 3$, two thresholds R_{MIN} and R_{MAX} are to be set. Let (x_s, y_s) denote the set of points in BIN_{TEX} such that, $BIN_{TEX}(x_s, y_s) = 1$ and $R_{MIN} < D_{SEL}(x_s, y_s) < R_{MAX}$, where

$$D_{SEL}(x_s, y_s) = \sqrt{(x_s - x_c)^2 + (y_s - y_c)^2} \tag{2.22}$$

(x_c, y_c) being the centre of the binary map BIN_{TEX}. Then the radius of the circular mask denoted as R_{MASK} is obtained as the median over the entire set $D_{SEL}(x_s, y_s)$, i.e.

$$R_{MASK} = MEDIAN(D_s), s \in \{1, 2, 3, ...N_s\} \tag{2.23}$$

where N_s is the number of pixels in BIN_{TEX} satisfying $R_{MIN} < D_{SEL}(x_s, y_s) < R_{MAX}$. Results of this adaptive circular mask generation process are shown in Fig. 2.14.

3

Hand-crafted feature selection and adaptation

Contents

(a) Duroc (b) Ghungroo (c) Hampshire (d) Yorkshire

Figure 3.1: Exemplar muzzle images from the four breeds.

Once the muzzle interior is isolated using the adaptive BALL-based segmentation algorithm discussed in Chapter 2, measurements can be taken on both texture and colour fronts. These measurements taken within the muzzle region are then pooled together to generate macro-statistics which eventually form a part of the visual descriptor. While choosing these macro statistics, it is important to ensure that numbers produced on a breed specific basis (i.e. on conditional grounds), remain largely distinct. Exemplar muzzle images from the four different breeds: Duroc, Ghungroo, Hampshire and Yorkshire are shown in Fig. 3.13(a-d). The colour and texture characteristic of the muzzle for the four breeds are listed in Table 3.1. The motivation for developing the macro-statistics on the colour and texture front comes from these characteristics listed in Table 3.1.

It is clear from Table. 3.1 that while examining breed-differences in a pairwise fashion, Ghungroo and Yorkshire turn out to be antipodes both with respect to colour and texture. However, because of the dual-colouration which prevails in the case of both Duroc and Hampshire pigs, the breed classification problem becomes tricky. In the following we discuss about the development of texture and colour attributes and how they are coalesced to form macro-statistics which can impart distinctiveness to each of the four breeds.

Table 3.1: Distinguishing visual parameters from the original cropped muzzle colour images (taken from male pigs) and their corresponding GSMs.

	Duroc	**Ghungroo**	**Hampshire**	**Yorkshire**
Colour profile	Either completely powdery-black or Dual-coloured (pink and powdery-black with a slight white-tuft)	Greyish black	Dual coloured (pink and grey)	Pink
Texture profile	Dot pattern density high (only over powdery-black patch)	Dot density high virtually throughout the muzzle	Dot pattern density moderate (only over greyish segment of the muzzle)	Dot pattern density uniformly low over the entire muzzle

3.1 Texture features

The breed characteristics mentioned in Table 3.1 provide the hint that the amount of textural details present on the muzzle surface may serve as a distinct representation for each of the four breeds. The contour, pore and hair follicles on the muzzle surface represent these textural details. These textural details do not follow any particular orientation or relative positioning with respect to each other on a breed specific basis. Only the amount of these details matter in generating breed specific statistics. Thus, there arises a need to highlight these details. Further quantization of this feature set will ensure robustness to image variations within a specific class. The following are some observations regarding this textural details:

- Sweat pores are present all over the muzzle surface both on the pinkish-white regions (present in breeds such as Hampshire, Duroc and all-pink Yorkshire), as well as the greyish-black patches (present in breeds such as blackish-Ghungroo, dual-coloured Hampshire and selectively dual-coloured Duroc).

- Hair or cilia (both prominent as well as stunted) are different for different breeds. Firstly these are confined to the greyish-black regions and do not exist over the pink patches. Over the greyish-black patches this density is more or less uniform and high. Hence, Ghungroo, which is all black, exhibits a high density of pores and hair (both stunted and

long) all around the dial, while in the case of Yorkshire, there are no hair/stunted hair on the muzzle surface.

The objective of any texture filter should be to produce distinct and different results/outputs for different breeds, while swallowing the variability within the same class. The textural details present on the muzzle surface need to be highlighted out at their corresponding location on the muzzle surface for this purpose. There are several texture descriptors available in literature [21]. However, in order to capture the amount of textural details on the muzzle surface two texture filters/descriptors have been used for trapping the significant portions in the muzzle region where there are artefacts related to sweat-pores, hair, stunted hair, patch transition regions (or contours) and in some cases wrinkles in the skin. The Gradient Significance Map [8] and the Morphological Top-hat(THAT) [1] operators are used to generate two separate binary maps which can then be quantized to produce what are known as patch density maps(PDMs) [8]. These two operators are discussed next.

3.1.1 Gradient Significance Map Generation

The structured artefacts present over a muzzle surface involve lines(from cilia/hair follicles), several scattered small circles and dots(stemming from the sweat pores) and internal contours if the muzzle is dual coloured as in the case of Hampshire and Duroc. If one deploys a conventional discrete differentiator, then the image would become too noisy and the dot density pattern would remain virtually uniform across all breeds. If Gaussian smoothing is combined with discrete differentiation, then the smoothing parameter σ associated with the Gaussian function can be carefully chosen so that the finer unwanted dots and thin lines are eliminated from the counting process. The details of this operation are explained with the help of the following steps:

Step-1: A Gaussian smoothed derivative along the X and Y directions are computed by controlling the smoothing parameter σ. The discrete derivative is a discretized version of the derivative of the zero centric Gaussian function,

$$\begin{aligned} D(x) &= \frac{d}{dx}\left[\frac{1}{\sqrt{2\pi\sigma^2}}\exp^{(-x^2/2\sigma^2)}\right] \\ &= -\frac{x}{\sqrt{2\pi\sigma^{3/2}}}\exp^{(-x^2/2\sigma^2)} \\ &= -c_0 x\exp^{(-x^2/2\sigma^2)}, x \in \mathbb{R} \end{aligned} \tag{3.1}$$

The value of standard deviation (or σ) plays a crucial role in shaping the gradient profile of the muzzle. Relatively small values of sigma would amplify sharp ciliary patterns, expiration pores and contour boundaries. For registered and cropped images of muzzles, this value of σ would correspond to one or two pixels in resolution (leading to a derivative footprint of approximately 6σ or 6 pixels to 12 pixels respectively). Large values of σ would tend to iron out the sharp details, ignoring the spot/spike introduced by the cilia/hair or even pores during the imaging process. To preserve the internal details it is therefore advisable to select a value of σ between one and four. The question lies as to how the upper bound on σ is decided. For a 512×512 image, the typical thickness of a cilia and the diameter of a sweat pore is around $10 - 12$ pixels as observed from Fig. 3.2. So long as the length 6σ is comparable with the dimensions of the structured artefacts, the lines and sweat pores will be preserved. All other micro-details and variations will be dissolved.

Step-2: Both the X and Y discrete kernels (K_x and K_y) are generated from $D(x)$ through sub-sampling. Thus,

$$\begin{aligned} K_x(x=0, y=i) = D(i), for i == -M, -(M-1), ..., 0, ..., (M-1), M \\ where, M = floor[3\sigma] K_x(x,y) = 0 elsewhere \\ and K_y(x,y) = K_x(y,x) \end{aligned} \tag{3.2}$$

Step-3: If C_{MUZZ} is one of the snout/muzzle color images as seen above and I_{MUZZ} the intensity pattern of the muzzle, the intensity profile is derived from the RGB palette as follows:

$$I_{MUZZ} = 0.3C_R + 0.587C_G + 0.114C_B \tag{3.3}$$

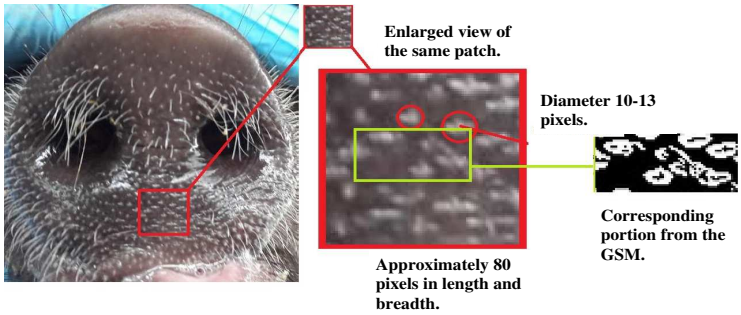


Figure 3.2: Figure showing the approximate dimensions of pores and cilia on the muzzle surface.

where, C_R, C_G and C_B represent the red, green and blue channel profiles of the muzzle color image.

Step-4: If $I_{MUZZ}(i,j)$ represents the intensity level at pixel location (i,j) ,the X and Y gradient profiles are computed at each pixel after the image is zero padded on all sides, as a two dimensional convolution process:

$$G_x(i,j) = \sum_{u=-M}^{M}\sum_{v=-M}^{M} K_x(u,v)I_{MUZZ}(i-u,j-v) = I_{MUZZ(pad)}(i,j) * K_x(i,j)$$
$$G_y(i,j) = \sum_{u=-M}^{M}\sum_{v=-M}^{M} K_y(u,v)I_{MUZZ}(i-u,j-v) = I_{MUZZ(pad)}(i,j) * K_y(i,j) \quad (3.4)$$

where the '*' operator here represents 2D discrete space convolution.

Step-5: The magnitude of the smoothed gradient is computed at each pixel as:

$$M_G(i,j) = \sqrt{G_x(i,j)^2 + G_y(i,j)^2} \quad (3.5)$$

Step-6: The mean over all the gradient values is computed as:

$$\mu_G = \frac{1}{N_1 \times N_2}\sum_i\sum_j M_G(i,j) \quad (3.6)$$

where, $N_1 \times N_2$ is the size of each snout image.

Step-7: Generate a normalized gradient profile by dividing gradient magnitude profile by the mean gradient μ_G

$$M_{G(norm)}(i,j) = \frac{M_G(i,j)}{\mu_G} \tag{3.7}$$

Step 8: A significance map is constructed by thresholding the normalized gradient profile. This significance map is a binary map and the threshold value δ_G, is chosen as UNITY. Any pixel which exhibits a normalized gradient value larger than ONE is set to ONE and the others are set to zero. The reason for this is because we are interested only in the points or locations where the gradient values which have some significance (hence the baseline threshold is chosen as the one because all the gradient values are normalized with respect to the mean). This ensures that the threshold remains content adaptive and rides with variations in contrast, illumination changes, variations induced because of erroneous de-focussing of the muzzle by the camera. This significance map is represented by a binary image of size $N_1 * N_2$,

$$BM_{GSM}(i,j) = \begin{cases} 1 & M_{G(norm)}(i,j) > \delta_G \\ 0 & otherwise \end{cases} \tag{3.8}$$

This binary map BM_{GSM} thus generated is termed as the Gradient Significance Map(GSM). Fig. 3.3, shows the muzzle profile of one of the pigs and the corresponding GSM.

If the threshold δ_G is too small, there will be an uncontrolled classification of pixels as significant ones and almost all binary images irrespective of their breed will appear the same as in Fig. 3.4(b). One the other hand if δ_G is too large, very little detail will be captured and once again all the density images will begin to appear the same, irrespective of their breed as in Fig. 3.4(d). To facilitate sufficient base feature separation by an accurate representation of the concentration of pores and cilia on the muzzle, this threshold is set to an in-between value '1' (Fig. 3.4(c)). This binary map which we term as the GSM thus generated serves as the primary feature map from which suitable statistics need to be extracted out for breed segregation.

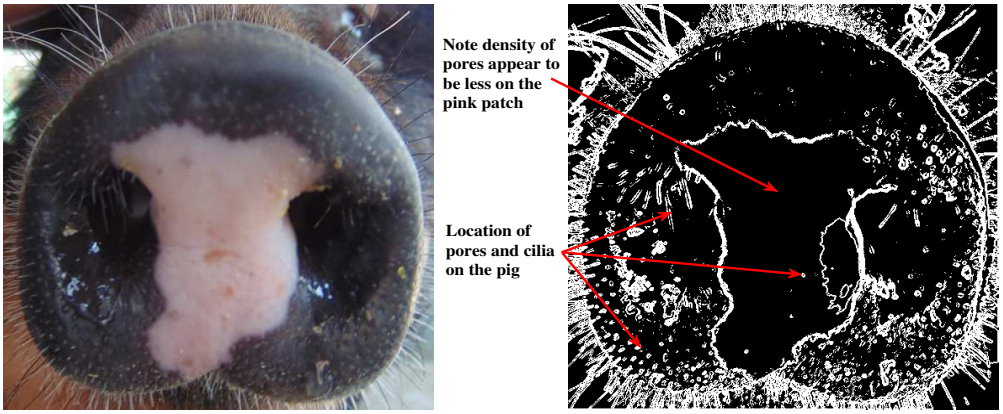


Figure 3.3: Muzzle image and binarized gradient profile.

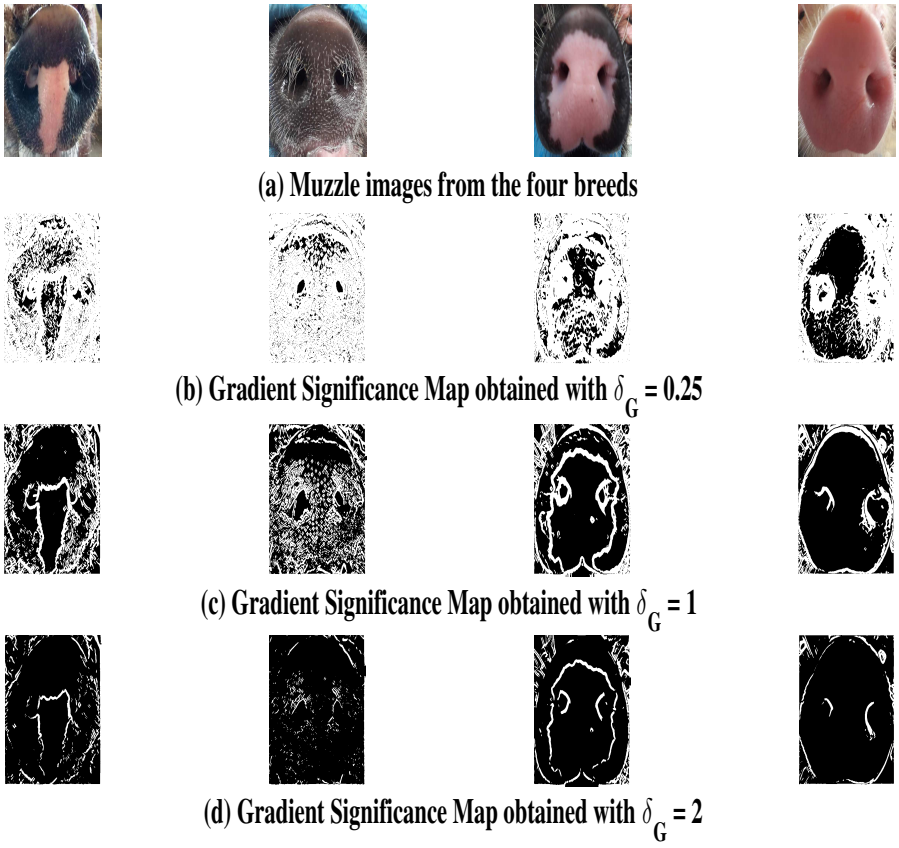
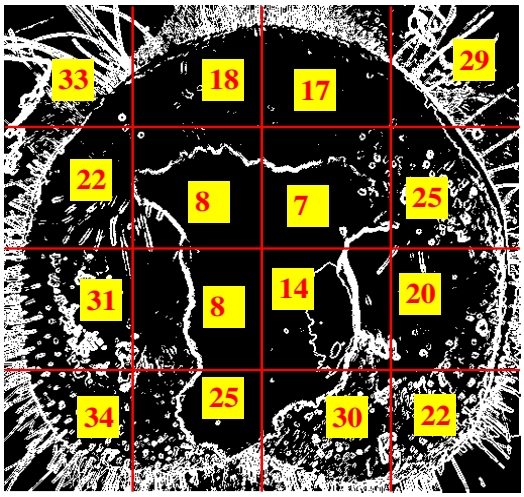


Figure 3.4: Effect of δ_G on the Gradient Significance Map.

3.1.2 Localized Texture Profiling and Maximum Likelihood Inferencing.

To demonstrate the efficacy of the GSM obtained using the Derivative of Gaussian operator as described in Section 3.1.1 towards breed classification, a formulation is shown where an attempt has been made to extract secondary statistics from the GSM based on a conjecture which is called as the patch diversity conjecture and use these statistics for classification. It is mentioned as follows:

Patch diversity conjecture: *The patch density profile is expected to be a function of the environment in which these animals are reared and is therefore expected to vary from breed to breed. More importantly the density maps derived from different spatial locations are expected to be different. It is precisely this diversity in these patch distributions that we wish to use for our final inferencing and decision making procedure related to breed identification.*



Pixel info: (301, 320) 1

Figure 3.5: Division of the GSM into patches and the corresponding patch statistic.

The distribution of textural patterns on the muzzle surface for the different breeds have been observed to follow a pattern. Ghungroo has uniformly high textural density all around the muzzle, whereas Yorkshire has uniformly low textural density throughout. For Hampshire, the textural density is uniformly low in locations of the pink patch and relatively high in other zones. In the case of Duroc, if the muzzle region has no pink patch, then the texture

is uniformly moderate all around; else for dual-coloured muzzle images, the textural density follows a similar pattern as Hampshire. To extract location specific statistics related to density of textural profile, the GSMs are first divided into equal sized patches as shown in Fig. 3.5. The statistic calculated for each patch is the percentage of significant pixels in that patch. Thus, if the patch size is $m \times n$ and n_s is the no. of significant pixels in that patch as obtained from the GSM, then the corresponding patch statistic as obtained from the GSM is

$$S(patch) = \frac{\sum_{(i,j)\in patch} BM_{GSM}(i,j)}{mn} \times 100\% \tag{3.9}$$

In order to verify the **Patch diversity conjecture** mentioned above, an experiment was carried out with a database of 311 muzzle prints corresponding to 30 animals across four breeds (Duroc, Ghungroo, Hampshire and Yorkshire). There were around 8 to 15 muzzle variations from each animal. Duroc and Ghungroo had six animals in their set while Hampshire and Yorkshire had 9 animals each. The muzzle images were taken with a high resolution hand-held camera with another person holding the nose of the pig tightly to avoid excessive blurring of the image due to relative motion between the camera and muzzle surface.

First the RGB colored image is converted to gray scale, resized to 1000×1000 and then the GSM was constructed. Square patches were prepared for analysis. Let N_P be the length associated with this square patch. Regarding the patch size N_P, if the patch size is too small, it will make the patch statistic too sensitive to camera panning and pig head movement (pose variation) and illumination changes, during image acquisition. On the other hand, large patch sizes are not desirable because variation of the textural density as a function of location on the muzzle surface is lost as a result. As a trade-off, we begin with a patch size of $N_P = 250$ resulting in a 4×4 grid with 16 patches. From these 16 patches, effectively 14 patches were analysed as the patches from the top-left and the top-right were left out of the analysis owing to background interference. One of the reasons for choosing a rectangular grid arrangement of patches was because the muzzle movement was insignificant and subsequently rotational effects and perspective differences were low. Thus the views were largely front on for a majority of the snapshots.

3.1.2.1 Training procedure

This training phase was set-up as follows:

- Using the proposed feature extraction algorithm the patch density statistics were computed for every muzzle image within the training set, which comprised of 159 muzzle prints (out of a total of 311 muzzle images) coming from 30 different animals across four breeds. The patch statistics were computed for a patch size of 25% (viz. the muzzle prints were split into 4×4 grids).

- Thus each print was mapped to a 4×4 matrix:

$$\mathbf{P}(\text{Image} - k) = \begin{pmatrix} S_{11}(k) & S_{12}(k) & S_{13}(k) & S_{14}(k) \\ S_{21}(k) & S_{22}(k) & S_{23}(k) & S_{24}(k) \\ S_{31}(k) & S_{32}(k) & S_{33}(k) & S_{34}(k) \\ S_{41}(k) & S_{42}(k) & S_{43}(k) & S_{44}(k) \end{pmatrix} \tag{3.10}$$

  with $S_{ij}(k) \in [0, 1]$ indicating the fraction of significant points in patch located at position (i, j) and $i, j \in \{1, 2, 3, 4\}$ corresponding to Pig-k.

- From the $N_T = 159$, training muzzle prints across four breeds, the 4×4 patch matrices for each breed were concatenated.

$$\begin{aligned} \mathbf{DUROC}_T &= \{\mathbf{D}_1, \mathbf{D}_2, ..., \mathbf{D}_{N_D}\} \\ \mathbf{GHUNG}_T &= \{\mathbf{G}_1, \mathbf{G}_2, ..., \mathbf{G}_{N_G}\} \\ \mathbf{HAMP}_T &= \{\mathbf{H}_1, \mathbf{H}_2, ..., \mathbf{H}_{N_H}\} \\ \mathbf{YORK}_T &= \{\mathbf{Y}_1, \mathbf{Y}_2, ..., \mathbf{Y}_{N_Y}\} \end{aligned} \tag{3.11}$$

  with, $N_D + N_G + N_H + N_Y = N_T = 159$.

- If $\mathbf{BR}_k$ corresponds to a patch matrix from PIG-k in breed type **BR**, this can be written

down as,

$$\mathbf{BR}_k = \begin{pmatrix} S_{11}(BR_k) & S_{12}(BR_k) & S_{13}(BR_k) & S_{14}(BR_k) \\ S_{21}(BR_k) & S_{22}(BR_k) & S_{23}(BR_k) & S_{24}(BR_k) \\ S_{31}(BR_k) & S_{32}(BR_k) & S_{33}(BR_k) & S_{34}(BR_k) \\ S_{41}(BR_k) & S_{42}(BR_k) & S_{43}(BR_k) & S_{44}(BR_k) \end{pmatrix} \tag{3.12}$$

where, $S_{ij}(BR_k) \in [0, 1]$. All the patch statistics from a breed corresponding to a spatial index (i, j) were concatenated to create a location specific conditional histogram. For instance the conditional histograms for patch location $(i, j) \in \{1, 2, 3, 4\}$ for the four breeds can be created by first sorting the values from the patch statistics (from a specific breed), in ascending order and then binning the count of the values falling within a fixed range. If M is the number of histogram bins, this process is represented as follows:

$$\begin{aligned} \hat{\mathbf{f}}_{S_{ij}/DUROC}(x) &= BIN_M\left[Sort\left(\{S_{ij}(D_1), S_{ij}(D_2), , ..., S_{ij}(D_{N_D})\}\right)\right] \\ \hat{\mathbf{f}}_{S_{ij}/GHUNG}(x) &= BIN_M\left[Sort\left(\{S_{ij}(G_1), S_{ij}(G_2), , ..., S_{ij}(G_{N_G})\}\right)\right] \\ \hat{\mathbf{f}}_{S_{ij}/HAMP}(x) &= BIN_M\left[Sort\left(\{S_{ij}(H_1), S_{ij}(H_2), , ..., S_{ij}(H_{N_H})\}\right)\right] \\ \hat{\mathbf{f}}_{S_{ij}/YORK}(x) &= BIN_M\left[Sort\left(\{S_{ij}(Y_1), S_{ij}(Y_2), , ..., S_{ij}(Y_{N_Y})\}\right)\right] \end{aligned}$$

where, $Sort(.)$ sorts the array of scalars in the ASCENDING order and $BIN_M(.)$ generates the fractional count of values in M equi-spaced bins over the range $[0, 1]$.

- To ensure there is some form of a polynomial fit for the histograms, the Gaussian density function, which has two degrees of freedom, has been chosen as a reference:

$$f_{S_{ij}/BR}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{3.13}$$

with, $BR \in \{$DUROC, GHUNG, HAMP, YORK$\}$. In the training phase, we learn the parameters of this Gaussian fit i.e μ, σ. A plot of all the learnt Gaussian distributions are shown in Fig. 3.6. Note each subfigure in the set (a to n) comprises of a parametric fit for each of the four histograms corresponding to a specific patch location (i, j) with patch locations $(1, 1)$ and $(1, 4)$ not considered on account of extreme background information

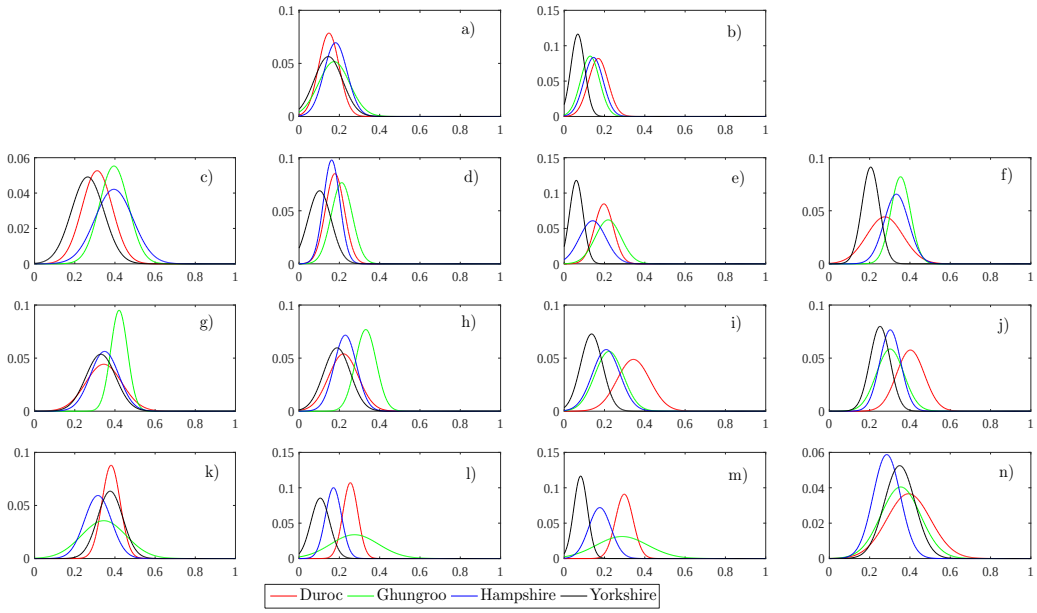


Figure 3.6: Patchwise conditional densities for each of the four breeds: .

and irrelevant details.

The following are some observations regarding the sub-figures:

- Sub-figures Fig. 3.6(a,b) still show considerable overlap in the conditional density functions on account of prevalent background information in all the patches extracted from these two spatial locations irrespective of the breed-type.

- A clear discrimination between the conditional densities begins with sub-figure(c) and continues all the way to sub-figure (n).

- As predicted by the patch density conjecture, since Yorkshire has been reared in colder areas, the fraction of significant points corresponding to pores and cilia is much smaller as compared to the other breeds (corroborated by Fig. 3.6(g,h)). This is depicted by the 'black' conditional density function in Fig. 3.6((g,h,k,n), which has the smallest mean in almost all the patches.

- On the other hand because of the high density of pores and cilia for Ghungroo, the

conditional mean is much higher than the other breeds for most of the patches (Green Gaussian curve in Fig. 3.6(d,e,f,g,h,l) and corroborated by Fig. 3.6(c,d)).

3.1.2.2 Testing and Inferencing procedure

When a query muzzle template is supplied to this spatial conditional patch distribution model, the statistics related to the density of textural details across different patches are first computed according to (3.9).

Thus, this query muzzle print becomes a 14-point vector:

$$\begin{aligned}\bar{Q} \quad = \quad & [q_{1,2}, q_{1,3}, q_{2,1}, q_{2,2}, q_{2,3}, q_{2,4}, q_{3,1}, \\ & q_{3,2}, q_{3,3}, q_{3,4}, q_{4,1}, q_{4,2}, q_{4,3}, q_{4,4}]\end{aligned} \tag{3.14}$$

The patchwise inferencing is done as follows: For each patch corresponding to the spatial location (i, j), the breed with which the corresponding query patch is most closely associated is extracted through a simple MAXIMAL LIKELIHOOD test.

$$\hat{BR}_Q(i,j) = \underset{BR}{ARG\,MAX} \left\{ \mathbf{f}_{S_{ij}/DUROC}(q_{ij}), \mathbf{f}_{S_{ij}/GHUNG}(q_{ij}), \mathbf{f}_{S_{ij}/HAMP}(q_{ij}), \mathbf{f}_{S_{ij}/YORK}(q_{ij}) \right\} \tag{3.15}$$

where, $BR \in \{DUROC, GHUNG, HAMP, YORK\}$. The overall association of the query vector with one of the breeds is obtained by taking a MAJORITY VOTE across all patch decisions:

$$BR_Q(FINAL) = MAJORITY_{VOTE} \left[\hat{BR}_Q(1,2), \hat{BR}_Q(1,3), \hat{BR}_Q(2,1), ..., \hat{BR}_Q(4,4) \right] \tag{3.16}$$

If $BR_Q(FINAL)$ is the same as the original breed, then the query has been IDENTIFIED correctly, otherwise there is a mis-classification. For testing: 29 muzzle prints from Duroc, 34 from Ghungroo, 45 from Hampshire and 44 from Yorkshire were deployed out of the total of 311 and the number of correct detections were noted in Table. 3.2. Since Yorkshire demonstrated a low conditional mean and a small variance across most patches, as expected the classification percentage was high (100%).

Table 3.2: Results of the breed classification algorithm with patch size $N_P = 250$

Breed	Classification Accuracy
Duroc	75.86%
Ghungroo	70.59%
Hampshire	57.78%
Yorkshire	100%

Table 3.3: Confusion matrix associated with the breed classification algorithm: patch size set as $N_P = 250$. Ideally the diagonal elements must be as close to '1' as possible.

	Duroc	Ghungroo	Hampshire	Yorkshire
Duroc	22/29	5/29	1/29	1/29
Ghungroo	4/34	24/34	6/34	0
Hampshire	4/45	7/45	26/45	8/45
Yorkshire	0	0	0	44/44

Duroc and Ghungroo showed moderate classification (slightly poorer) results of 75% and 70% respectively, as the variances in their conditional density functions were larger leading to a significant overlap in the functions. Since Duroc, Ghungroo and Hampshire all have partial white patches in some pigs, Duroc tends to be confused for Ghungroo and Hampshire (Ghungroo higher, because of the similarity in the contour and overall muzzle structure) and Yorshire (least), while, Ghungroo tends to be confused for Duroc and Hampshire (both high) and Yorkshire (least). This can be witnessed in the confusion matrix Table. 3.3.

Hampshire shows the worst classification result of 58% as the location and size of the internal pink patch is arbitrary as shown in Fig. 3.7.Since the pink patch location and sizes are arbitrary, misclassification as either Ghungroo or Yorkshire is equally possible as is also evident from the confusion matrix in Table 3.3. This result also justifies the need for a colour descriptor.

Thus the secondary statistics derived from the GSM are reasonably effective in segregating the breeds. Apart from Hampshire, the classification accuracy for the other three breeds have crossed 70% mark which is a fairly encouraging result obtained using a single type of texture operator. Even for Hampshire, the classification accuracy is more than 50% which is better than a random guess. In the following, one more texture descriptor based on the Morphological

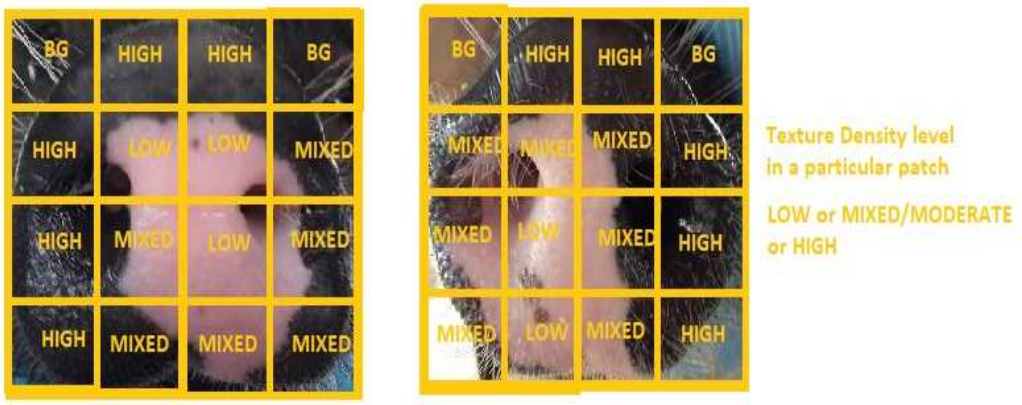


Figure 3.7: Texture density variation patch-wise for two different Hampshire pigs.

Tophat operator is discussed followed by the colour features.

3.1.3 Morphological Top-hat operator

The morphological Top-hat transformation ($THAT$) operator is applied on the gray scale muzzle image and the response is then thresholded to obtain another binary map similar to the GSM. The morphological Top-hat transformation operator is defined as

$$\hat{f} = f - (f \circ b) \tag{3.17}$$

where f is the gray scale muzzle image, b is the structuring element and $(f \circ b)$ denotes the morphological opening operation of f by b. The response $T_{hat}(f)$ generated in (3.17) is then thresholded with respect to a preset threshold TH to generate a binary map BM_{THAT}, i.e.

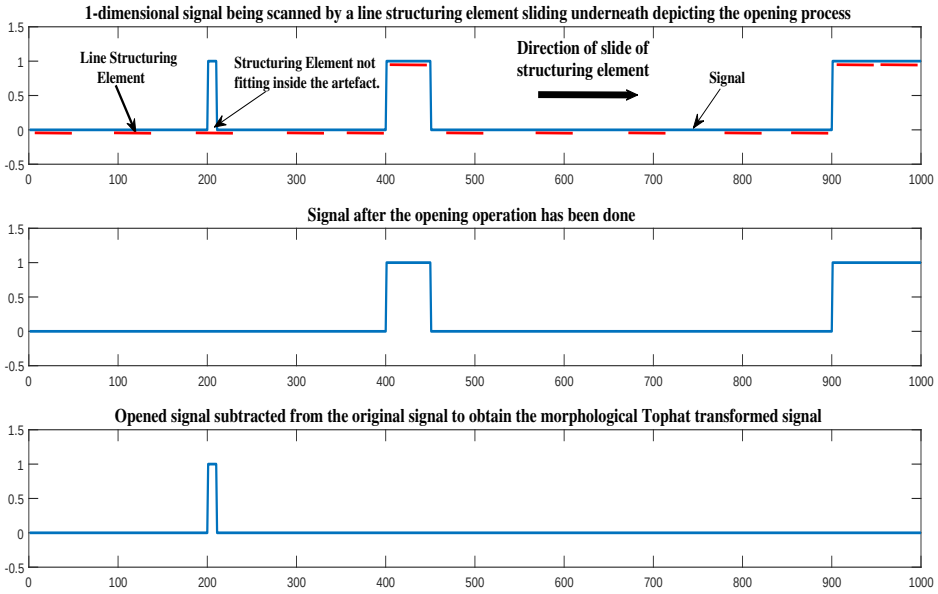


Figure 3.8: Illustration of the morphological Top-hat transformation process [1]. Here, the morphological top-hat transformation is being carried out on a signal in $\mathbb{R}^1$ with a line structuring element.The top figure shows the structuring element sliding underneath the signal to perform the morphological opening operation. The middle figure shows the signal after the opening operation has been done on it. The bottom figure is the difference between the original signal and the opened signal to obtain the Top-hat transformed signal.

$$BM_{THAT}(i,j) = \begin{cases} 1 & \hat{f}(i,j) > TH \\ 0 & otherwise \end{cases} \tag{3.18}$$

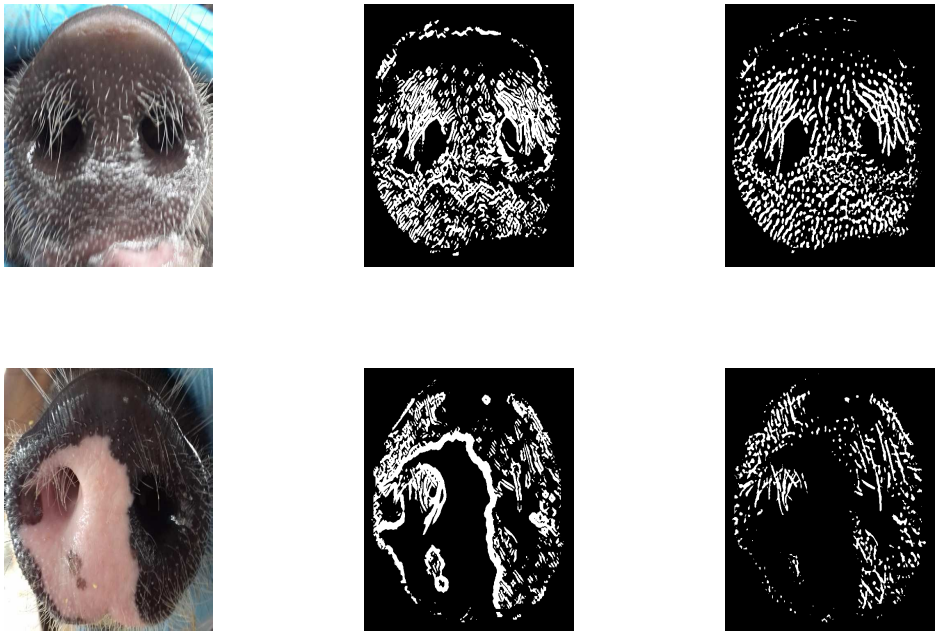
The morphological Top-hat transformation operator is used to highlight bright objects in dark background. Most of the textural details viz. the pores and hair follicles are distributed in the greyish region of the muzzle, which make them appear like bright objects embedded in a relatively darker background. This provides the motivation to use this operator, so that hair follicles and pores are exclusively picked up rejecting irrelevant details like the contour separating the grey region from the pink in the case of certain breeds such as Duroc and Hampshire.

The process of working of the morphological Top-hat transformation operator is shown in Fig. 3.8. In the case of intensity images, the morphological top-hat operator is expected to

highlight discontinuities having finite width at least along one of the direction. This includes structures such as points, roof edges and line edges. The hair follicles and pores on the muzzle surface fall under these category of structures. The dimensions of the structuring element is crucial in this regard. The structuring element should be larger in dimension as compared to the object to be highlighted. This operator when applied on the muzzle images highlight only the hair follicles and pores on the surface. This property is particularly useful for segregating Ghungroo from the other three breeds. As discussed earlier Ghungroo has the highest amount of hair follicles and pores among all the four breeds and so our interest lies in highlighting these details only in the binary map. If the Derivative of Gaussian operator is applied, the contour of the muzzle boundary along with the line separating the pink and grey region (for Hampshires and Duroc) also gets manifested as foreground pixel in the GSM. This acts as a hindrance in the segregation of Ghungroo from the rest. The binary maps BM_{GSM} and BM_{THAT} generated using the Derivative of Gaussian and the Morphological Top-hat operator are shown in Fig. 3.9. The muzzle image in the top row corresponds to a Ghungroo muzzle image while the muzzle image in the bottom row corresponds to a Hampshire muzzle image. By comparing columns (b) and (c) for the Hampshire muzzle image in the second row it can be clearly seen that the boundary separating the grey and pink region gives a strong response in BM_{GSM} when the Derivative of Gaussian operator is used. However, this boundary does not appear in the binary map BM_{THAT}.

3.1.4 Arriving at a Stable Texture Descriptor for each Muzzle Image.

The classification acccuracies obtained for the four breeds using the Patch Diversity Conjecture mentioned in Section 3.1.2 were relatively low for all the breeds except Yorkshire, and for Hampshire in particular it was extremely low. This is because of the inherent flaw in the Patch Diversity Conjecture which assumes that the density maps at different spatial locations are expected to be different for all the breeds. However, this conjecture is highly invalid for dual coloured muzzle like Hampshire, where the location and size of the pink patch is arbitrary as



(a) (b) (c)

Figure 3.9: Illustration of the details extracted out in BM_{GSM} and BM_{THAT}. Column (a) shows two different muzzle images. Column (b) shows their corresponding binary map BM_{GSM} extracted using the Derivative of Gaussian operator. Column (c) shows their corresponding binary map BM_{THAT} extracted using the Morphological Top-hat operator.

shown with the help of the example in Fig. 3.7. Thus, it becomes necessary to arrive at a more stable texture descriptor which is robust to these variation of textural density with location on the muzzle surface. The binary maps BM_{GSM} and BM_{THAT} are significance maps which can then be quantized to produce what are known as patch density maps(PDMs) [8]. A binary significance map (either BM_{GSM} or BM_{THAT}), is split into four equal quadrants and the fraction of white/significant pixels are counted and recorded in each quadrant. Fig. 3.10(a) shows the muzzle of a Hampshire pig and Fig. 3.10(b,c) shows the differential information obtained from the significance maps with respect to BM_{GSM} and BM_{THAT} along with the quadrant PDM scores. The contour present in the interior, because of the pink patch in the Hampshire pig, is picked up by the Derivative of Gaussian operator but not by the Morphological Top-hat operator. This distinction becomes useful while segregating certain breeds with respect to others in the texture domain.

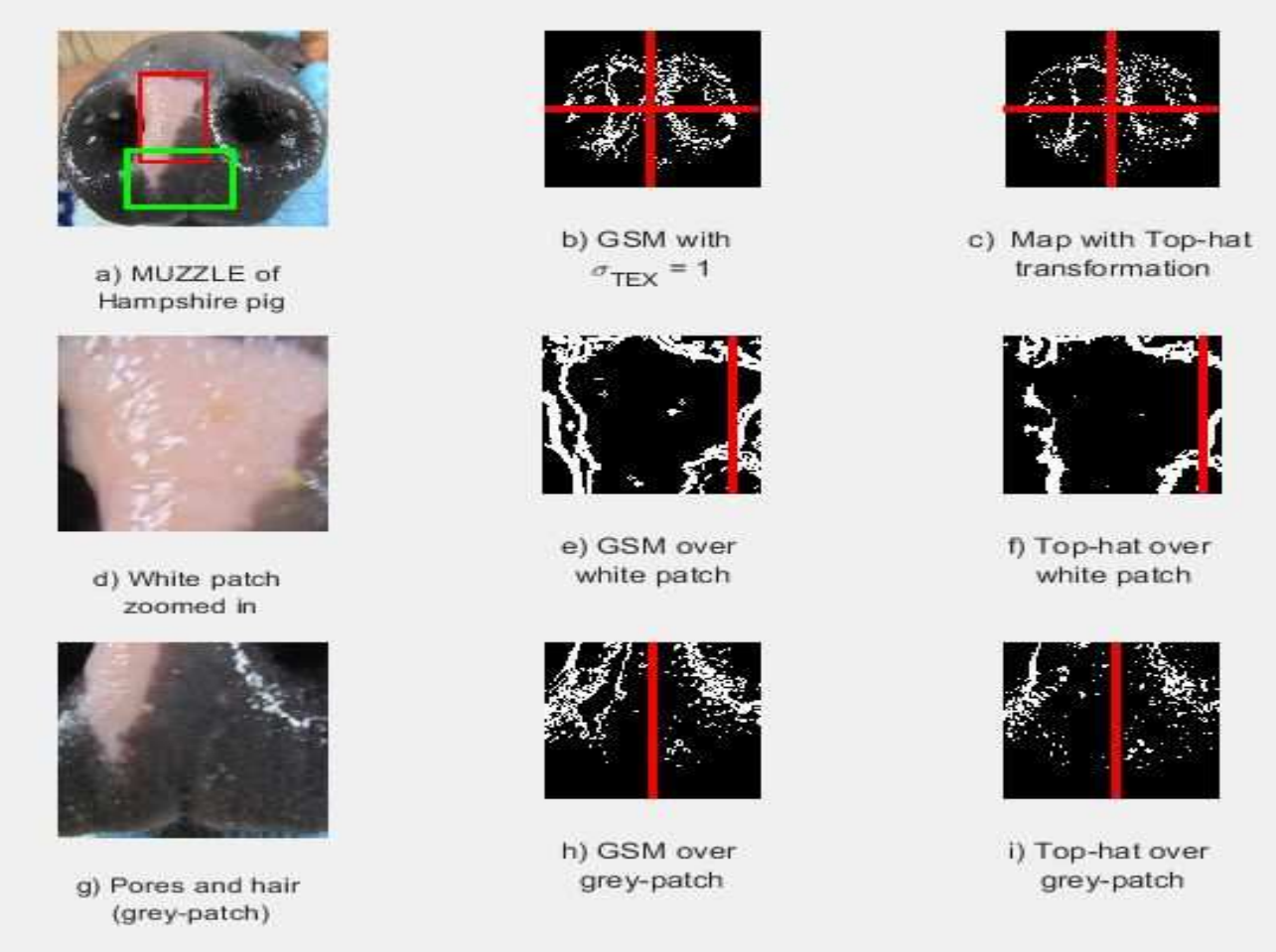


Figure 3.10: Patch related texture variations seen in a Hampshire pig.

The significance map BM_{GSM} is a function of the standard deviation parameter (σ_{TEX}) associated with the Derivative of Gaussian operator, while the Morphological Top-hat operator is primarily a function of the radius of the structuring element r. It has been observed [8] that a small value of σ_{TEX} tends to enhance the noise leaking into the feature calculation, while a large value suppresses key internal details on the muzzle's surface. For the morphological Top-hat operator, the dimension of the structuring element must be carefully selected [1]. The main objective of using this operator is to enhance fine details like the hair follicles and pores on the muzzle surface, while filtering out the coarse details. Hence, a structuring element with a small dimension is desirable. The corresponding quadrant-wise density feature vectors, related to the binary maps BM_{GSM} and BM_{THAT} are respectively given by,

$$\begin{aligned} Q_{GSM}(\sigma_{TEX}) &= [p_1, p_2, p_3, p_4] \\ Q_{THAT}(r) &= [q_1, q_2, q_3, q_4] \end{aligned} \tag{3.19}$$

where, $p_i, q_i \in [0, 1]$ and $i \in \{1, 2, 3, 4\}$. With reference to the size of the patches described by the parameter N_P in Section 3.1.2, it is to be noted that N_P has been set to 50% of the image dimensions for evaluating the density feature vectors Q_{GSM} and Q_{THAT} in contrast to N_P being set as 25% of image dimensions in Section 3.1.2. This increase in the patch size has been done to mitigate the variability, particularly in dual coloured Hampshire and Duroc pigs. In the case of the Hampshire pig, owing to the presence of the pink-patch, there is a prominent internal contour which is picked up in BM_{GSM} but suppressed in BM_{THAT}. Thus the texture filter outputs are different for Hampshire (Fig. 3.11). The lack of details within the pinkish-white patch in the Hampshire pig is demonstrated in Fig. 3.10(d,e,f). Much of the details (dot, blob, line and curve artifacts) are concentrated over the greyish-black zone, Fig. 3.10(g,h,i). To

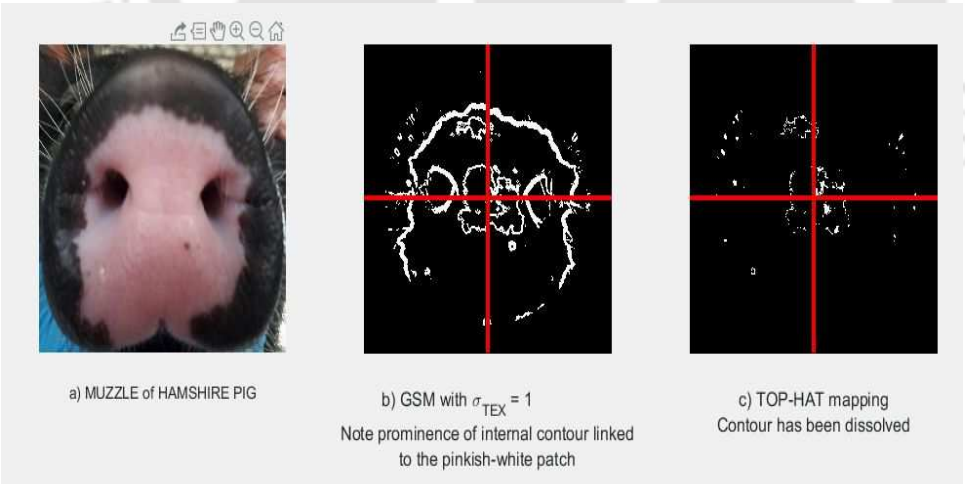


Figure 3.11: Texture profile and patch density scores for a Hampshire pig.

account for the overall density scores over the four quadrants, two mean parameters have been derived from the BM_{GSM} and BM_{THAT} quadrant density scores.

$$\begin{aligned} \mu_{GSM} &= \frac{p_1 + p_2 + p_3 + p_4}{4} \\ \mu_{THAT} &= \frac{q_1 + q_2 + q_3 + q_4}{4} \end{aligned} \tag{3.20}$$

The final texture feature vector is a 10-dimensional vector, a function of two primary parameters: σ_{TEX} from the GSM and structural elemental radius r for the top-hat operator.

$$f_{TEX}(\sigma_{TEX}, r) = [p_1, p_2, p_3, p_4, q_1, q_2, q_3, q_4, \mu_{GSM}, \mu_{THAT}] \tag{3.21}$$

With respect to the attributes in the texture feature vector $f_{TEX}(\sigma_{TEX}, r)$, the following trends are anticipated in the case of the four breeds: For Yorkshire, all the attributes are expected to have low values, while for Ghungroo all of them are expected to be high. For Hampshire, any two or three out of four values are expected to be low and means are expected to be moderate. In the case of Duroc, if the muzzle is dual coloured the attributes are expected to follow the same trend as Hampshire, while for Duroc muzzle images which are not dual coloured, all the attributes including the means are expected to have moderate values. To verify the robustness and distinctiveness characteristic of this feature vector, muzzle images of 20 pigs (five per breed), were chosen for the experiment. The four breeds were Duroc, Ghungroo, Hampshire and Yorkshire. The Gaussian gradient parameter σ_{TEX} was varied over the range of $\sigma_{TEX} \in \{2, 4, 6, 8, 10\}$ and the radius of the top-hat structuring element was varied as $r \in \{2, 3, 5, 7, 9\}$.

In order to assess the performance of the texture feature as a function of the smoothing parameter from the Derivative of Gaussian (σ_{TEX}) and radius of the structuring element(r) from the Morphological Top-hat operator, the overall separability across breeds for different parameter settings was computed. This was done using a metric based on the Mahalanobis distance [37] for measuring cluster/class separability. Five muzzle images selected from five different pigs per breed were chosen for this purpose; leading to four different sets of feature vectors. Let BR_1,BR_2, BR_3 and BR_4 represent the four breeds/clusters. Then $d_{MD(i)}(j)$ is chosen to be the Mahalanobis distance from the centroid of the cluster BR_i to cluster BR_j. First, the distance of breed BR_1 from each of BR_2, BR_3 and BR_4 is measured individually. The overall distance of BR_1 (denoted by $\bar{d}_{MD(1)}$) from the other three classes is the mean of

Table 3.4: Separation scores for the composite texture feature for different values of σ_{TEX} (GSM-operator) and different radii r(THAT [1] morphological operator).

	$r = 2$	$r = 3$	$r = 5$	$r = 7$	$r = 9$
$\boldsymbol{\sigma_{TEX} = 2}$	31.50	45.04	39.33	35.57	35.77
$\boldsymbol{\sigma_{TEX} = 4}$	43.77	**52.41**	41.41	35.94	35.65
$\boldsymbol{\sigma_{TEX} = 6}$	44.30	51.77	41.18	35.43	36.68
$\boldsymbol{\sigma_{TEX} = 8}$	41.33	49.59	40.47	34.80	37.08
$\boldsymbol{\sigma_{TEX} = 10}$	36.73	46.37	37.92	32.27	34.78

the individual distances of BR_1 from the other three classes, i.e.,

$$\bar{d}_{MD(1)} = (d_{MD(1)}(2) + d_{MD(1)}(3) + d_{MD(1)}(4))/3 \tag{3.22}$$

In a similar manner, distances $\bar{d}_{MD(2)}$, $\bar{d}_{MD(3)}$ and $\bar{d}_{MD(4)}$ are measured. The overall separation between the four breed datasets is the mean over those four scores.

$$SM(\sigma_{TEX}, r) = \frac{\bar{d}_{MD(1)} + \bar{d}_{MD(2)} + \bar{d}_{MD(3)} + \bar{d}_{MD(4)}}{4} \tag{3.23}$$

From the separation scores produced in Table. 3.4, $\sigma_{TEX} = 4$ and $r = 3$ are obtained as the best parameter set for extraction of robust texture features. The corresponding t-SNE [38] map which brings out the feature separability across breeds for the optimal parameter set $\sigma_{TEX} = 4$ and $r = 3$ from Table. 3.4 is shown in Fig. 3.12.

3.2 Colour features

As mentioned earlier in Chapter 1, the motivation for the use of colour descriptors come from the observation that there is some distinctiveness in the colour profile between the various breeds. While Ghungroo muzzle images are fully greyish black in colour, Yorkshire muzzle images are fully pink in colour. Hampshire muzzle images are dual coloured. Some of the Duroc muzzle images have a colour profile similar to that of Hampshire, while others have a powdery black colour. Also, it has been observed that the texture descriptors are not sufficient to segregate the breeds. For example, some of the Hampshire muzzle images have a very large pink region with the muzzle, because of which the texture profile appears similar to Yorkshire.

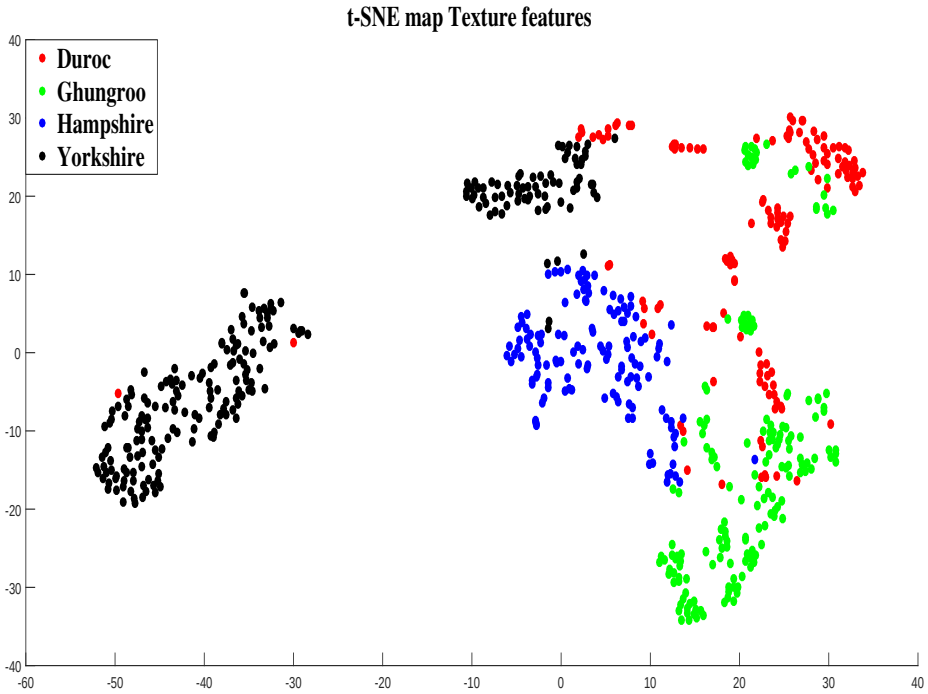
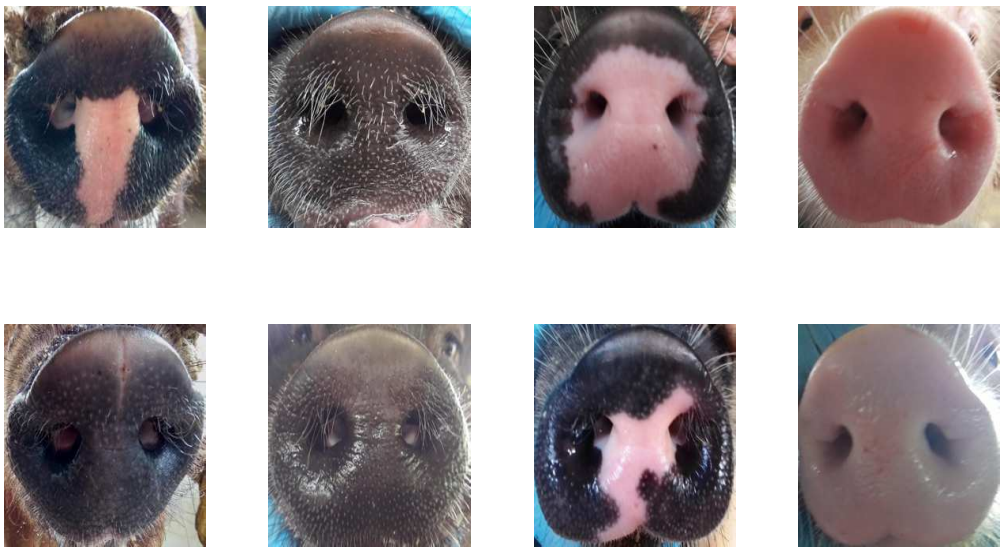


Figure 3.12: t-SNE map of the texture feature set.

On the other hand, Hampshire also shares similar texture profile with some of the Duroc pigs. In a very few cases, some of the Duroc pigs have texture profile similar to that of Ghungroo. Thus, there arises the need for colour descriptors to increase the separability between the breeds in feature space for such cases. Depending on the positioning of the light source and the manner in which the pig's snout is being held, there will be significant illumination variations across pigs from the same breed. This results in considerable intra-class variability. To ensure a robust colourimetric analysis, the luminance component must be segregated from the chrominance part and thus the image must be first converted from RGB space to another appropriate colour space. The luminance component gets decoupled from the chromatic part when the image is analyzed in the YCbCr space, making the Chromatic components (C_b and C_r) independent of the local illumination profile and variations as shown in [39]. Some muzzle samples from the four breeds (corresponding to $4 \times 2 = 8$ different pigs: two per breed), are reproduced here in Fig.3.13 for convenience, with the photographs taken under natural lighting conditions (viz.

in a shed in broad daylight). The development of the proposed colour descriptor based on a 2-dimensional histogram over the chrominance space (or $C_b - C_r$-space), is discussed.



(a) Duroc (b) Ghungroo (c) Hampshire (d) Yorkshire

Figure 3.13: Exemplar muzzle images from the four breeds.

3.2.1 Primary Colour Feature Map Obtained from $C_b - C_r$ histogram

It was summarized in Table. 3.1, that the colour composition of the muzzle surface shows some distinctiveness for the four breeds: Duroc, Ghungroo, Hampshire and Yorkshire. Examples of intra-class variability as well as breed separability with respect to colour can be observed in Fig. 3.13. In some cases, the muzzle of Duroc carries a pinkish-white patch (whose size and position is un-predictable). Hampshire male pigs on the other hand show a strong and consistent pink-patch presence (again the size and position of the patch is variable, but generally found much larger than the ones found in the odd Duroc pig). Yorkshire pig-muzzles are completely pink in colour, which makes them easily identifiable and separable from pure-greyish black Ghungroo pigs. The main confusion thus arises between Duroc and Hampshire both with respect to colour and texture. In the colourimetric part of the $Y - C_b - C_r$ space, this separation can be tapped statistically via $C_b - C_r$ histograms (Fig. 3.14 (third and fourth rows)). It is evident from row-4 (top-view of the $C_b - C_r$ histograms), that the chromatic-centroids of Duroc,

Ghungroo, Hampshire and Yorkshire are all distinct (Fig. 3.14, fourth row). The footprints of the histograms in the (a-b) space are much larger for Duroc, Hampshire and Yorkshire as compared to Ghungroo. The colour diversity is much less in the case of Ghungroo and maximum for Duroc (Fig. 3.14, fourth row). Thus, there are distinct features in this histogram for all the four breeds, which qualifies this primary feature as a robust yet distinctive colour descriptor for breed-segregation.

Colour measurements were taken from the muzzle surface only over the region defined by the mask. Thus, given a muzzle image, the following dataset associated with chroma measurements in $C_b - C_r$ space, was generated from the region defined by the mask: $CDATA_{PIG,BR} = \{(u_1, v_1), (u_2, v_2), ...(u_n, v_n)\}$, with n being the number of data-points picked picked from this region. The muzzle images were resized to 512×512 and all points within the circular mask were considered for the 2D histogram generation and colourimetric analysis. Robust statistics from this histogram such as, marginal and joint moments, secondary statistics such as Sarle's bimodality coefficient [22], both at a 1-D level and also at a 2-D level with the assumption of independence were computed. All the moments were normalized with respect to the standard deviations in both the dimensions (C_b, C_r).

Let $f_H(\tilde{u}, \tilde{v})$ denote the 2-dimensional normalized histogram associated with the chromatic-components (C_b, C_r) linked to the muzzle-colour profile of a particular pig belonging to a certain breed BR, where $BR \in \{1, 2, 3, 4\}$. This in essence is a 4-class classification problem where class-1 corresponds to the Duroc breed of pigs, class-2 to Ghungroo, class-3 to Hampshire and class-4 to Yorkshire. Here the variables $\tilde{u}$ and $\tilde{v}$ takes the C_b and C_r values respectively of a particular pixel. The ordered pair $(\tilde{u}, \tilde{v}) \in Z_S \times Z_S$, where $Z_S = \{-128, -127, ..., 128\}$. If $(u_i, v_i), i = 1, 2, ..., n$ denote the (C_b, C_r) values of the n randomly selected pixels in the region defined by the circular mask, then $f_H(\tilde{u}, \tilde{v})$ is defined as the total number of pixels having a quantized $(C_b - C_r)$ pairing of $(\tilde{u}, \tilde{v})$

$$f_H(\tilde{u}, \tilde{v}) = \sum_{i=1}^{n} \delta_{2D}(\tilde{u} - u_i, \tilde{v} - v_i) \tag{3.24}$$

Since $f_H(\tilde{u}, \tilde{v})$ is a 2-dimensional histogram, hence it has the following properties:

$$0 \leq \left(\frac{1}{n}\right) f_H(\tilde{u}, \tilde{v}) \leq 1$$
$$\left(\frac{1}{n}\right) \sum_{\tilde{u}=-128}^{128} \sum_{\tilde{v}=-128}^{128} f_H(\tilde{u}, \tilde{v}) = 1 \tag{3.25}$$

3.2.2 Secondary Features Extracted from the $C_b - C_r$ histogram

The statistics computed on this sub-sampled dataset tend to characterize the histograms seen in Fig. 3.14. Of interest are the following parameters: (P1) Bi-modality Index [22]: In breeds like Hampshire and Duroc, muzzles tend to show a pink patch and the rest of the muzzle is either greyish-black or powdery-black. Thus the histograms in these two cases tend to be of a bi-modal nature(if a pink patch is indeed present in the Duroc pig). This bi-modality, reflects as a heavy-tailed distribution and hence as per the literature [22], can be trapped using a combination of the Skewness and Kurtosis. Extension to 2-dimensional data is done here based on certain assumptions; (P2) Centroid: Mean vector associated with the chroma-pair (C_b, C_r). Since the muzzle is black for Ghungroo, weighted combination of black and pink for Hampshire, pink for Yorkshire and selective weighting (black,pink) for Duroc, hence the centroids are definitely expected to be distinct for these breeds; (P3) Footprint of the distribution: This is obtained through a principal component analysis (PCA), over the chroma-space. The eigen-vectors of the covariance matrix obtained from the data gives the principal components and the corresponding eigen-values represent the variance of the data along each of these components. Thus the square-root of the product of the eigenvalues is expected to provide an estimate of the footprint of the distribution; (P4) Skewness and Kurtosis of both the C_b and C_r data (to be used for computing the composite bi-modal index assuming independence of the colour-channels). The corresponding equations for parameters P1, P2, P3 and P4 are constructed in the following way:

First the marginal means of the respective chromatic components C_b and C_r are computed to generate the centroid of the distribution:

$$\mu_{C_b} = \frac{1}{n}\sum_{i=1}^{n} u_i$$
$$\mu_{C_r} = \frac{1}{n}\sum_{i=1}^{n} v_i \tag{3.26}$$

Then the Kurtosis and Skewness measures for the respective chromatic components C_b and C_r were computed as,

$$SKEW_{C_b} = \frac{\frac{1}{n}\sum_{i=1}^{n}(u_i - \mu_{C_b})^3}{(\sqrt{\frac{1}{n}\sum_{i=1}^{n}(u_i - \mu_{C_b})^2})^3}$$
$$KURT_{C_b} = \frac{\frac{1}{n}\sum_{i=1}^{n}(u_i - \mu_{C_b})^4}{(\frac{1}{n}\sum_{i=1}^{n}(u_i - \mu_{C_b})^2)^2} \tag{3.27}$$

In a similar fashion $SKEW_{C_r}$ and $KURT_{C_r}$ are computed. Based on Sarle's proposition [40], bi-modality can be predicted using the heavy tailed nature of the respective chroma marginal probability distributions, which in turn reflects in the higher order moments: Skewness and Kurtosis. This Bi-modality index can be computed for the respective chromatic components C_b and C_r as,

$$BIM_{C_b} = \frac{1 + SKEW_{C_b}^2}{KURT_{C_b}} \tag{3.28}$$

with a similar form for the C_r-component, BIM_{C_r}. Assuming independence of the chromatic components C_b and C_r, a 2D bi-modality coefficient approximation can be generated as a geometric mean of the bi-modality coefficients of C_b and C_r,

$$BIM_{APP(2D)} = \sqrt{BIM_{C_b} \times BIM_{C_r}} \tag{3.29}$$

To establish the size of the footprint of the joint $C_b - C_r$ probability distribution, a PCA analysis is done to compute the eigenvalues λ_1 and λ_2. The eigenvectors are discarded as they are expected to be data-sensitive but the eigenvalues are retained. Let the data-vector be $\bar{v}_i = [u_i, v_i]^T$ and the centroid $\bar{v}_{CEN} = [\mu_{C_b}, \mu_{C_r}]^T$. First the colour covariance matrix over the

$C_b - C_r$ space is computed as:

$$S_{C_bC_r} = \frac{1}{n}\sum_{i=1}^{n}(\bar{v}_i - \bar{v}_{CEN})(\bar{v}_i - \bar{v}_{CEN})^T \tag{3.30}$$

Let the eigen-decomposition of this matrix be,

$$S_{C_bC_r} = \mathbf{V}D\mathbf{V}^H \tag{3.31}$$

where, D is a diagonal eigenvalue matrix comprising of two eigenvalues: λ_1 and λ_2. The footprint of the joint probability distribution can be approximated as,

$$FOOT_{C_bC_r} = \sqrt{\lambda_1 \times \lambda_2} \tag{3.32}$$

The skew associated with this footprint, about the eigen-directions can be quantified as,

$$FOOT_{SKEW} = \frac{MIN(\lambda_1, \lambda_2)}{MAX(\lambda_1, \lambda_2)} \tag{3.33}$$

The final colour descriptor or feature vector is given by this seven-dimensional vector: $f_{COLOR} = [cd_1, cd_2, cd_3, cd_4, cd_5, cd_6, cd_7]$. Here, $cd_1 = \mu_{C_b}$, $cd_2 = \mu_{C_r}$, $cd_3 = BIM_{C_b}$, $cd_4 = BIM_{C_r}$, $cd_5 = BIM_{APP(2D)}$, $cd_6 = FOOT_{C_bC_r}$, and $cd_7 = FOOT_{SKEW}$. The t-SNE map [38], which is indication of the extent of feature separability across these four breeds is shown in Fig. 3.15. It can be observed from the t-SNE map that most of the feature vectors from Ghungroo are well separated out from the others and form a distinct cluster. There are however some Ghungroo muzzle images which have a colour profile very similar to that of Duroc, due to some of the muzzle images from both the breeds having common shades of grey. Duroc also has some similarity with Hampshire, due to some of Duroc pigs having dual coloured muzzle like that of Hampshire. On the other hand, it can be seen that there is some similarity between Hampshire and Yorkshire muzzle images as well. This is because, some of the Hampshire muzzle images have a significantly large pink area within the muzzle region which makes it appear very similar to Yorkshire. Also in some of the Yorkshire muzzle images, due to heavy specular reflection, the entire pink region is not properly picked up/represented by the colour features.

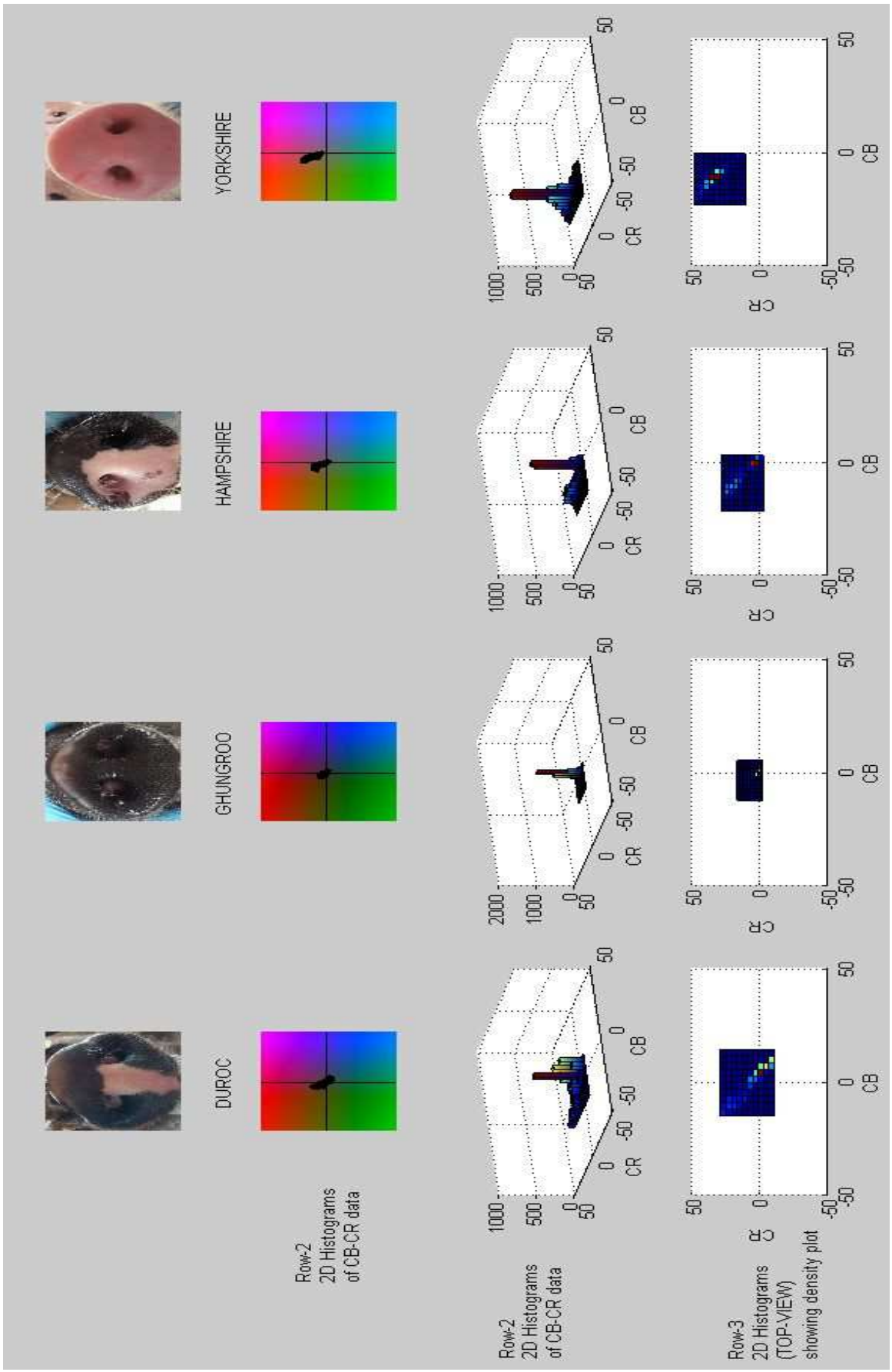


Figure 3.14: Scatter plots and 2D histograms in the Cb-Cr domain (four pig breeds).

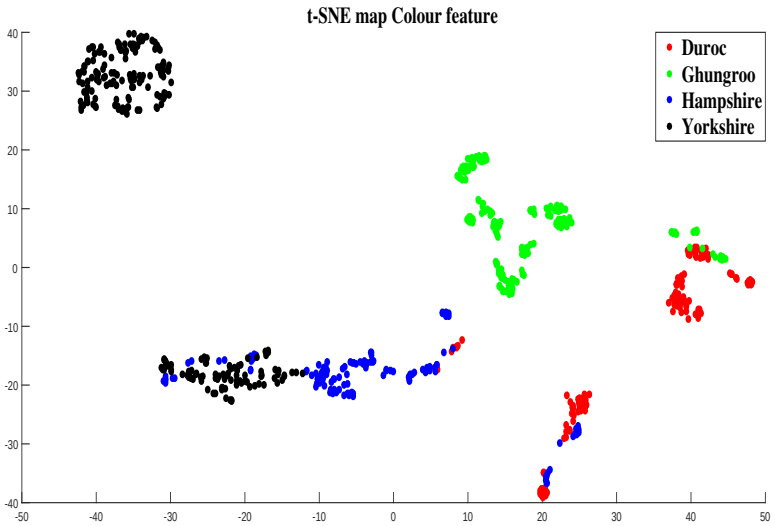


Figure 3.15: 2-dimensional t-SNE map for colour.

4

Graph Synthesis for Hierarchical Classification

Contents

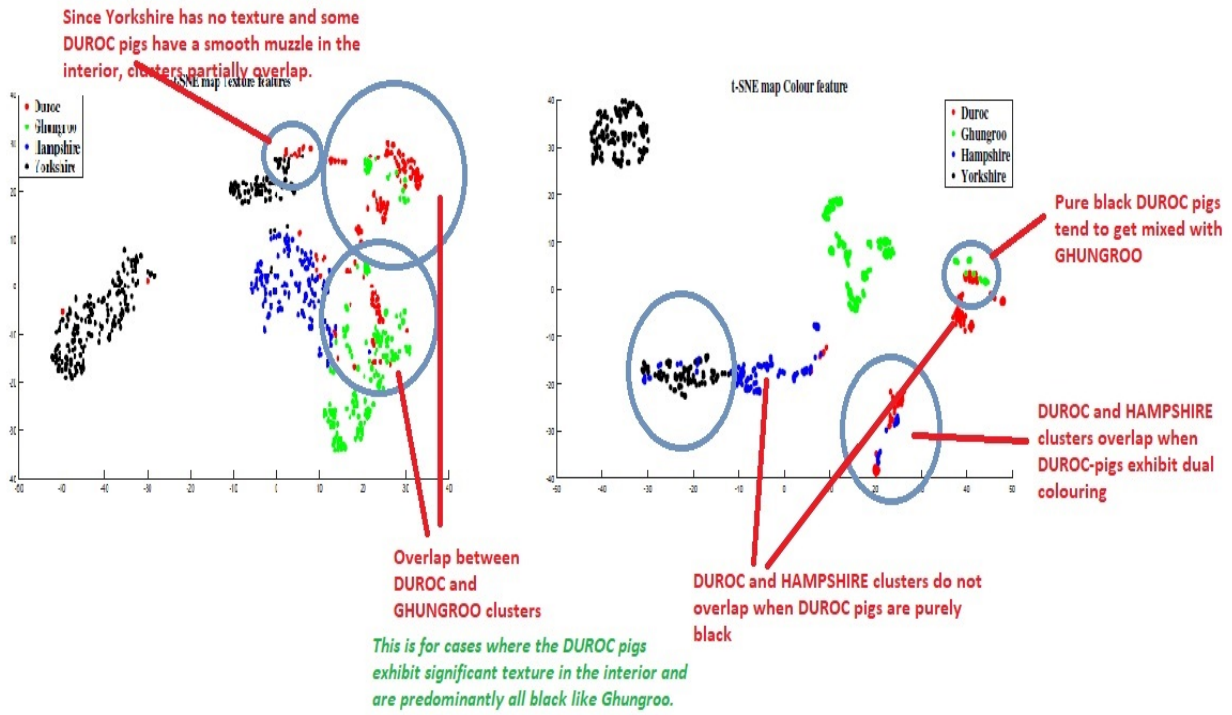


Figure 4.1: Figure highlighting the separability between the breeds in the colour and texture domain.

The feature vectors extracted from the colour and texture descriptors are not strong enough individually to provide segregation between all the breeds. As mentioned earlier in Table 1.2, the different breeds cannot be separated from each other using a single feature type and this can be verified from the t-SNE plots of colour and texture features in Fig. 4.1. From this figure, it can be observed that Duroc and Ghungroo have considerable overlap in the texture domain, however, they are better separated in the colour domain. Ghungroo is well separated from Hampshire in colour domain, but in texture domain, the two clusters are close to each other. Also, because a very few of the Duroc pigs have a smooth muzzle, they overlap with Yorkshire in the texture domain but they are far apart in the colour domain. Thus a single feature-type is not sufficient to segregate all the breeds in feature space.

Fig. 4.2 identifies specific patterns of feature vector distribution in feature space due to which standard classifiers like SVM, nearest neighbour algorithms are expected to perform poorly. In the texture feature space, Duroc and Ghungroo cannot be separated with SVM classifier because the feature vectors from Duroc are highly scattered in feature space and there

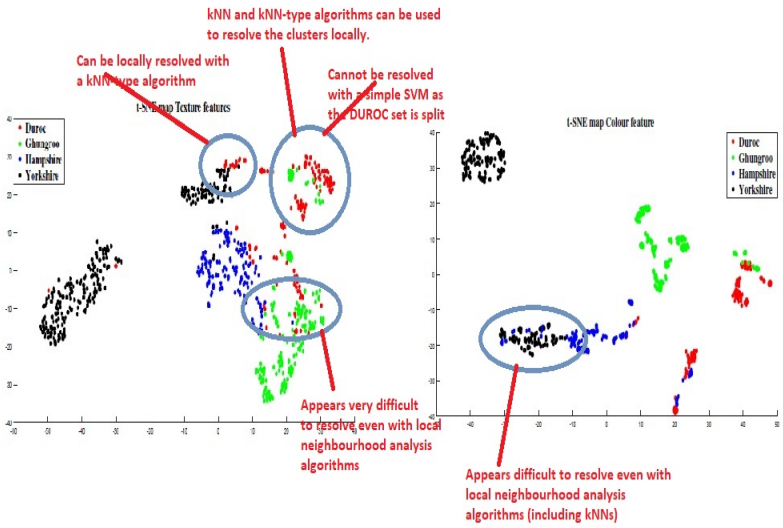


Figure 4.2: Figure showing certain patterns in the distribution of feature vectors in feature space which can affect the performance of standard classifiers.

is considerable overlap between the two classes. Even nearest neighbour algorithms are also not expected to perform well under such cases. This same situation is observed in the colour feature space between the two breeds Duroc and Hampshire. Thus it is not possible to separate out all the breeds in any single feature space using standard classifiers like SVM or the class of nearest neighbour algorithms. The combination of colour and texture features increases the separability between breeds only if there is strong correlation between the two feature types. As will be shown later in this chapter that this correlation is weak for breeds such as Duroc and Ghungroo, hence the combination of colour and texture features is not expected to improve the separability between breed-pairs which involve Duroc or Ghungroo. Thus, it can be stated that it is not possible to separate out the breeds in a single classification stage using Colour(C), Texture(T) or a combination of Colour and Texture($C \cup T$) features. This provides the motivation to go for a multi-stage classification scheme, where the breeds are siphoned out one after another in a hierarchical fashion. The breeds which are better separated from the rest are siphoned out first to prevent classification errors from propagating down the hierarchy to the later classification stages. Furthermore, at each stage of hierarchy the decision space which best separates the two classes is used. For example, Ghungroo seems to be best separated from the others in the

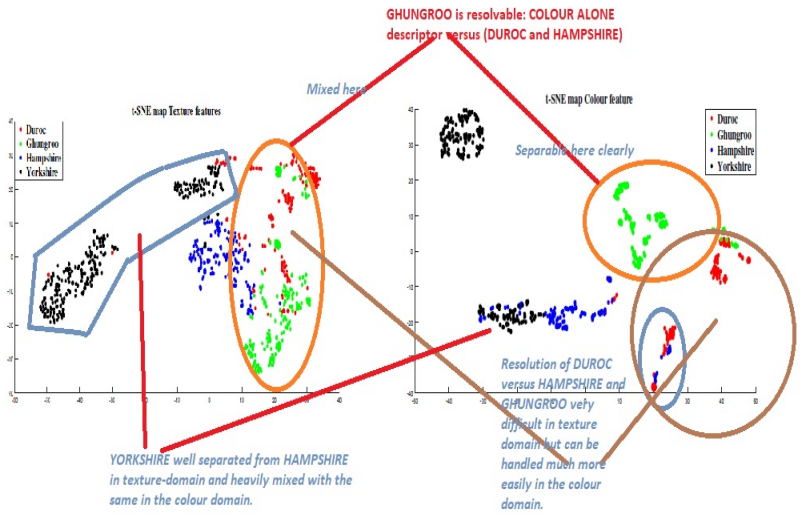


Figure 4.3: Motivation for choice of separate feature space for each breed.

C feature space as shown in Fig. 4.3. It is to be noted that at each stage of classification, a standard classifier like SVM can be used. The choice of the hierarchy as well as the decision space for each classification stage is decided by our algorithm which will be described in the subsequent sections.

As seen from the earlier figures and the initial discussion, feature vectors obtained from the colour or texture information alone are not self-sufficient for breed discrimination. On the other hand, simply combining the texture and colour features into one single composite vector may not work on all comparison fronts. Given a particular breed such as Ghungroo, it was spotted in Fig. 4.3, that resolution on the texture alone front or combination of texture and colour front is not possible. However, it can be easily resolved on the colour alone front with respect to anyone of the other three breeds. In some cases such as Hampshire, which is dual coloured and the texture profile varies in accordance with the size of the pink patch(as the pink zones in all pink and dual coloured muzzle carry no hair and exhibit very few apparent sweat pores in the texture map), the choice of feature can be done based on which select group of breeds is weighed against it. For example, in the case of Hampshire versus Duroc, both colour and texture seem to be the right choice; in the case of Hampshire versus Yorkshire the choice could again be colour and texture as the dual colouration in Hampshire can be captured by the

Sarle's index. In the case of Hampshire versus Ghungroo, they are better separated by colour. This is the motivation for taking a graph-theoretic approach mainly for identifying the order and manner in which the breeds are siphoned out in the proposed classification procedure. A decision tree is constructed whose leaves form the breed-nodes but the difference here is that unlike a conventional decision tree algorithm which does random attribute sampling and model building and adaptation to arrive at the optimal decision space at each node, the feature selection in our proposed tree building algorithm is done at the MACRO level and NOT at the attribute level.

4.1 Tree based Classification Methods in Literature

Since the pig breed classification task calls for a tree like hierarchical classification scheme, hence it becomes necessary to take a look at other tree based classification algorithms in literature and compare their similarities and differences with our tree generation algorithm. In the following, a couple of well established tree based classification schemes are discussed.

4.1.1 Tree construction based on Phylogenetic Analysis

The concept of constructing a Phylogenetic tree [24] based on the similarity and differences in the physical or genetic characteristics between different species to describe the evolutionary process, can also be used for arriving at the optimal classification hierarchy. Given required information regarding breeds, the methods for the construction of this optimal phylogenetic tree can be classified mainly into three categories: (i) Distance based methods [41]: These methods are used when pairwise distances between the different entities are available for the construction of the phylogenetic tree. Each leaf node on the phylogenetic tree represents one entity. A hierarchical clustering algorithm is used for preserving the relative distance between different entities on the tree; (ii) Maximum parsimony [42]: This method searches for the phylogenetic tree with the minimum number of evolutionary steps which can explain a given set of data assigned on the leaves. Here, the topology of the tree is randomly changed till there is no more improvement in the parsimony; (iii) Likelihood-based methods [43]: This method

involves computing the likelihood of the given data sequence with standard evolution models and the tree corresponding to the best likelihood model is generated.

The requirement in all cases is to identify a tree which can maximize the separation between the child nodes at each step of evolution, so that the best classification accuracy is possible with the given set of features. Distance based methods using hierarchical clustering algorithms [44] are the best choice since they try to separate out the maximally distant classes at each step of evolution starting from the root node. Thus in the first stage the class which is farthest from all the other classes is separated out; in the second stage, out of the remaining classes, the one which is farthest from the remaining others is separated out and so on. The Bioinformatics Toolbox in MATLAB [25] provides functions related to Phylogenetic analysis using Distance based methods. The function takes the matrix of pairwise distances and uses a Hierarchical Agglomerative Nesting algorithm (AGNES) [44] to cluster objects based on their similarity. The algorithm starts by treating each object (breed) as a singleton cluster. Next, pairs of clusters are successively merged until all clusters have been merged into one big cluster containing all objects. The result is a tree-based representation of the objects, named *Dendrogram*, where the leaf-nodes correspond to the breed-types. At each step of the algorithm, the two clusters that are the most similar are combined into a new bigger cluster (nodes). This procedure is iterated until all points are member of just one single big cluster (root).

The process does not however prescribe a procedure for arriving at the right choice of features/statistics at each decision point (i.e. is feature agnostic) and also handling multiple features. If the matrix of pairwise distances is created by selecting the branch weight between any two breeds as the maximum distance over all the feature combinations for that breed-pair, then a tree can be constructed (even this feature is not available with the AGNES). The final phylogenetic tree which is built on this collection of maximal distances, remains therefore completely *feature type agnostic*. The AGNES therefore demands some form of a higher level protocol, first, to generate a distance table based on some fusion measure and then to identify the best feature at various levels in the tree.

4.1.2 Decision Trees

Decision trees [45] though being very simple machine learning techniques have the powerful property of being able to automatically select the attributes from the feature vector which can impart maximum separation between two or more classes of data given the training samples. Starting from the root node, each parent node is split into multiple child nodes in a way that the purity of each child node is maximized. The most widely used splitting criterion are based on Gini's Diversity Index [46] and entropy reduction methods [47]; although more recent methods are available in literature for binary [48] and non-binary splits [49] which ensure the constructed tree is more compact with smaller number of nodes without compromising on the classification accuracy. The major steps involved in the construction of a decision tree include: (i) Selecting one of the attributes(a_i) from the feature vector; (ii) Choosing a threshold(t_i) for that attribute, that divides the training data into child nodes; (iii) Measuring purity of the child nodes, when the parent node is split based on the attribute a_i using the threshold t_i (iv) Repeating this process for all the attributes and feasible t_i till maximum purity is obtained, for all the child nodes (v) On obtaining the maximum purity split, the process is repeated for a second split, and so on. The following are some issues which can be anticipated with this decision tree approach:

- Attribute selection is not the same as feature-type selection, as the selection is done by randomly sampling the parent composite feature set. Since this random attribute selection design is not in tune with the on-field analysis and customization, results are definitely expected to be poorer as compared to the optimal hierarchical tree generation algorithm (which includes feature-type identification).

- While decision trees are expected to work well with raw data vectors and simple primary statistics, performance will degrade when the feature vectors include secondary and robust statistics, which are compact in nature. Decimation of such secondary statistics is expected to result in an information loss both with respect to within-class similarity (which brings breed-specific variants together) with respect to that parameter (which has

been dropped), as well as precious information which imparts segregation across breeds. For instance dropping the bi-modality index will make Hamshire pigs look like Yorkshire (with respect to colour).

4.2 Proposed Hierarchical Classification Scheme

The proposed hierarchical classification scheme starts with computing a distance metric between clusters corresponding to all possible breed pairs in across all the feature spaces. The different feature/decision space in our case are formed by the Colour(C), Texture(T) and a combination of Colour and Texture ($C \cup T$) feature sets. Starting with four breeds and the pairwise distance between the breeds across different feature spaces, the proposed algorithm tries to find out a tree like structure; where at each node of the tree a particular breed is separated out from the rest. The algorithm also provides the decision space in which a particular breed can be separated out from the rest in the best possible manner. The order in which the breeds are separated out one after another is also decided by the proposed algorithm and this order in the hierarchy starts from the top of the tree down to the bottom. The proposed classification scheme is described in detail in the following subsections.

4.2.1 Cluster/Class separation indicators

The distance metric between any two classes should properly portray the separability between these two classes assuming that this frame is modelled using a binary linear classifier (two class separation problem). First a linear discriminant classifier is learnt from the data taken from the two classes which attempts to separate the two classes with minimum error. There are two independent indicators for this cluster separation:

Fractional crossovers: Here,the number of crossovers on either side of the hyperplane is used as a measure of separability. More the fractional number of crossovers, lower is the separability between the two classes (and greater is the class mixing). Let n_i and n_j denote the number of members in class i and class j respectively and n_{ij} the number actually belonging to class i

but falling on the other side of the hyperplane. This distance indicator is,

$$SEP_{CO}(i,j) = 0.5 \times \left(2 - \frac{n_{ij}}{n_i} - \frac{n_{ji}}{n_j}\right) \tag{4.1}$$

A value of $SEP_{CO}(i,j) \approx 1$ indicates that the classes are clearly separable (via a linear classifier).

Normalized cluster distance: Here the separation between the clusters is of greater concern as compared to the overlap between them. Non-overlapping closely positioned clusters are penalized as compared to more separated ones. A hyperplane is first learnt for separating the two classes using a linear discriminant classifier. A certain fraction of data is chosen from each of the two classes, which are nearest to the learnt hyperplane. Note that all of these chosen data points from either of the classes are the ones which fall on the respective side of the hyperplane to which they actually belong to. Let $\mu_{D(i)}(\alpha)$ and $\mu_{D(j)}(\alpha)$ denote the mean distance of these α-fraction of the nearest data-points of class i and class j respectively from the hyperplane. Then, distance metric along the second dimension is defined as

$$\begin{aligned} SEP_D(i,j) &= \left(\frac{S(\alpha)}{1+S(\alpha)}\right) \\ S(\alpha) &= \frac{\mu_{D(i)}(\alpha)}{\sigma_{D(i)}(\alpha)} + \frac{\mu_{D(j)}(\alpha)}{\sigma_{D(j)}(\alpha)} \end{aligned} \tag{4.2}$$

The role of standard deviation based normalization as far as the Euclidean distances are concerned is to penalize clusters which are non-compact (same centroidal separation, but show greater variability). The separation indicator is,

$$SEP_{OVERALL}(i,j) = SEP_{CO}(i,j) + SEP_D(i,j) \tag{4.3}$$

over range [0,2]. Table 4.2 gives the mean separation between two different breeds as a function of the feature/composite feature used over 100 different random selections of training data as will be explained later in Section. 4.3. It can be observed from Table 4.2 that out of the six breed-pairs, there are instances when either the color (C) or texture feature (T) alone gives

Table 4.1: Histogram type for colour and conditional distribution for texture density along with the correlation between (C) and (T) feature sets for various breeds.

Breed type	Histogram type for Color	Conditional distribution for Texture density	Correlation between Color and Texture feature parameters
Duroc	Color histogram can be either bi-modal or uni-modal(if uni-modal the spread is narrow).	Moderate to low texture density and moderate variability.	Virtually independent.
Ghungroo	Limited variability as far as color (grey shades) are concerned.	High density in texture but variability on the higher side.	Weak correlation (virtual independence of color and texture features).
Hampshire	Bi-modal.	Bi-modal.	Moderate to strong.
Yorkshire	Moderate variability in pink shades.	Low density in texture and limited variability (non-existent in most Yorkshire pigs).	Strong correlation

better class separability as compared to the union of colour and texture features $(C \cup T)$. Thus in the union of colour and texture features, the union may take place constructively or destructively, so that the class separation either increases or decreases respectively. Upon examining a combination of feature sets such as (C) and (T), the correlation profiles between these two feature sets in different breeds are expected to be different. Thus there is expected to be a mismatch between the correlation trends which affects the cluster separation in this common decision space$(C \cup T)$. There is therefore a need to exercise a judgement as to whether both features have to be included and if not, which one is the preferred choice to form the decision space.

Since the correlation between the feature sets (C) and (T) is strong both for Hampshire and Yorkshire as depicted in Table 4.1, hence the distance metric is observed to be maximum for the feature set $(C \cup T)$ as expected. Also, in all distance comparison between pairs of breeds involving Duroc or Ghungroo, because of the lack of correlation between the feature sets (C) and (T); it has been observed that the feature set $(C \cup T)$ does not provide the maximum distance between clusters with the exception of the Duroc-Yorkshire pair.

4.2.2 Guided Tree Selection

At a macro level, since a hierarchical classification strategy is to be adopted, based on the manner in which the training data is split and/or fused, for four breeds BR_i, BR_j, BR_k and BR_l, there are four distinct decision tree possibilities as shown in Fig. 4.4. The structure of the tree, which gives best classification results through a particular siphoning order at different levels and with a proper feature choice (or choices) at those decision points, needs to be identified without going through the computational rigor. Given four breeds, Duroc (D), Ghungroo (G), Hampshire (H) and Yorkshire (Y) (which will eventually become four leaf nodes in the final decision tree), there are six pairwise breed-cluster distances that can be computed: (D-G), (D-H), (D-Y), (G-H), (G-Y) and (H-Y). There are three macro feature possibilities: (i) Colour feature alone (C); (ii) Texture feature alone (including GSM as well as Top-hat) (T); (iii) Union of Colour and Texture (Composite) ($T \cup C$). A sufficient set of distances which can be used to derive a strategy for breed-siphoning is the $6 \times 3 = 18$-cell table of breed-cluster pairwise-distances shown in Table 4.2. From this Table secondary distances can be derived. To find out roughly how far each breed is cumulatively far away from the rest (with the same consistent rule), by pivoting around a specific breed, its distance from the other breeds are added up. This is in turn a function of the feature-type: Colour (C) or Texture (T) or Composite: Colour and Texture $C \cup T$. The following analysis will show that using the pairwise distances as a function of feature type as listed in Table 4.2, along with the analysis based on cumulative statistics leads to a classification tree of type T3(asymmetric binary tree). That is why the other tree structures in Fig. 4.4 have not been considered.

Table 4.2: Distance metric for all possible binary splits when two breeds are present: D:Duroc, G:Ghungroo, H:Hampshire, Y:Yorkshire; C: Colour features, T: Texture features, $T \cup C$: Composite features; Largest distances in each column are indicated in BOLD font.

	D-G	D-H	D-Y	G-H	G-Y	H-Y
C: Colour	**1.90**	1.71	1.74	**1.94**	**1.97**	1.64
T: Texture	1.67	**1.83**	1.81	1.81	1.95	1.87
$C \cup T$	1.71	1.74	**1.93**	1.87	1.94	**1.87**
Best features	C	T	$C \cup T$	C	C	$C \cup T$
MAX-distances	1.90	1.83	1.93	1.94	1.97	1.87

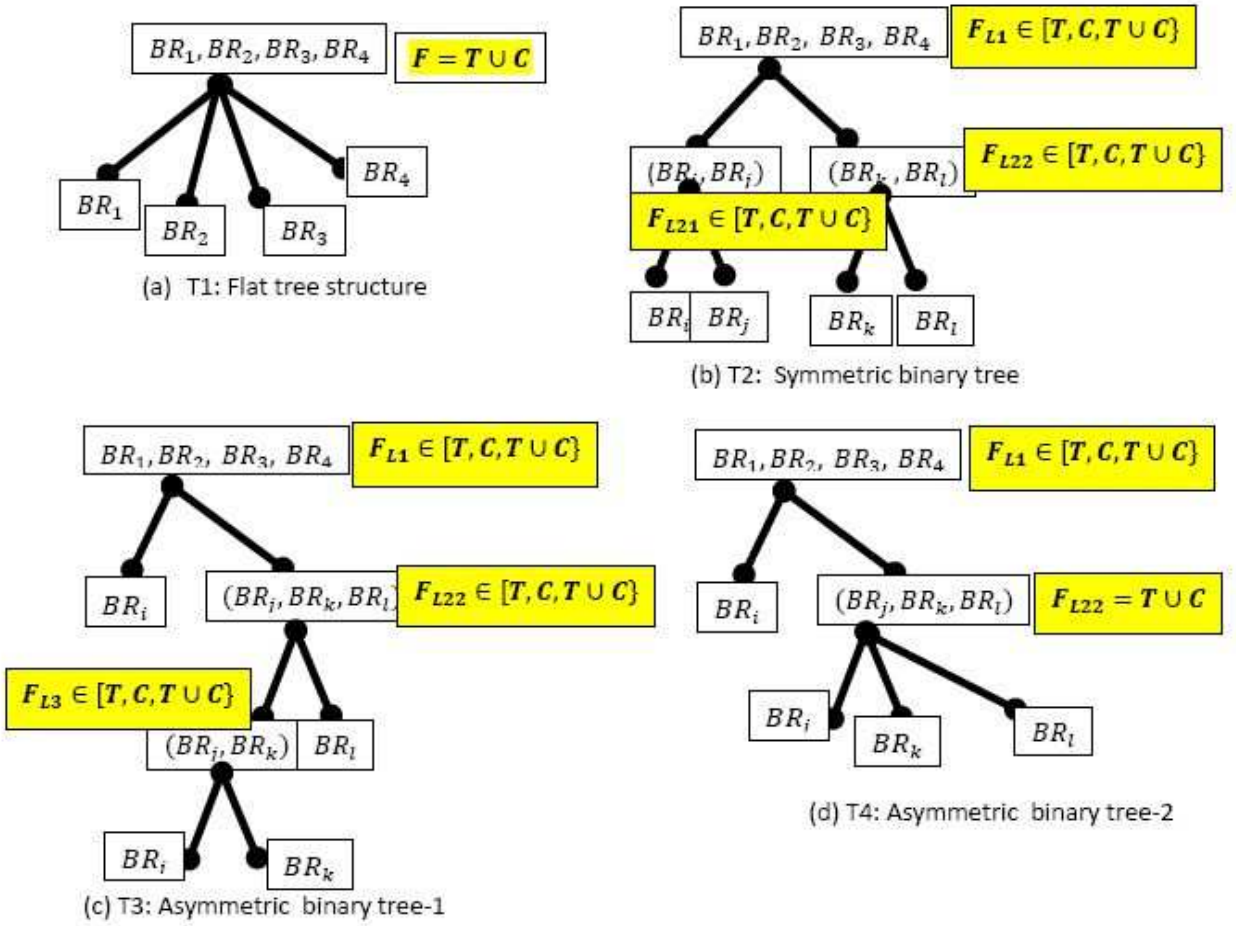


Figure 4.4: Classification routes or tree-types possible in a four class setting.

Table 4.3: Cumulative distances with respect to a particular breed across various feature combinations.

BREED and feature	Colour	TEXTURE	Colour and TEXTURE
Duroc vs rest	5.35	5.31	5.38
Ghungroo vs rest	**5.81**	5.43	5.52
Hampshire vs rest	5.29	5.51	5.48
Yorkshire vs rest	5.35	5.63	5.74
CUMULATIVE-MAX	5.81	5.63	5.74
CUMULATIVE-MEDIAN	5.32	5.47	5.50
DIFF(MAX, MEDIAN)	**0.49**	0.16	0.24

The cumulative distance table (Table. 4.3) provides a distinct angular perspective from the point of view of the individual breeds. The separation of Duroc (D) from the remaining breeds is a function of the feature combination used: Turns out to be $S_{DUROC}(C) = 1.90+1.71+1.74 = 5.35$ with respect to colour; $S_{DUROC}(T) = 1.67 + 1.83 + 1.81 = 5.31$, with respect to texture and $S_{DUROC}(C,T) = 1.71+1.74+1.93 = 5.38$, with respect to the composite feature involving colour and texture. Similarly such statistics can be computed for the other three breeds leading to Table. 4.3. For a specific feature type (color or texture or UNION), if the deviation between the maximum cumulative distance and median is significant, this indicates a certain skew in the breed arrangement and also provides indirect information, that the optimal tree structure, may not be a flat tree (i.e. Fig. 4.4(a)). Here, from the scores, Table. 4.3, appears to indicate that Ghungroo should be siphoned out first, in terms of color and then the cumulative distance table should be recomputed without Ghungroo. If one compares the differential (MAX, MEDIAN) scores for color (0.49), texture (0.16) and UNION (0.24) from Table. 4.3, there is a hint that the color feature could be dominant over texture, for this specific 4-breed arrangement.

Table 4.4: Leaving out Ghungroo, pairwise distances for various feature combinations, reproduced.

BREED and feature	D-H	D-Y	H-Y
C: Colour	1.71	1.74	1.64
T: Texture	**1.83**	1.81	1.87
$C \cup T$	1.74	**1.93**	**1.87**
Best features	T	$C \cup T$	$C \cup T$
MAX-distances	1.83	1.93	1.87

Table 4.5: Leaving out Ghungroo, cumulative distances with respect to a particular breed across various feature combinations.

BREED and feature	Colour	TEXTURE	Colour and TEXTURE
Duroc vs rest	3.45	3.64	3.67
Hamshire vs rest	3.35	3.70	3.61
Yorkshire vs rest	3.38	3.68	**3.80**
CUMULATIVE-MAX	3.45	3.70	3.80
CUMULATIVE-MEDIAN	3.38	3.68	3.67
DIFF(MAX, MEDIAN)	0.07	0.02	0.13

It is clear from the cumulative scores from Table. 4.3 that best results can be obtained from asymmetric binary split-tree types T3 (Fig. 4.4(c)) or T4 (Fig. 4.4(d)), where the solitary leaf

Table 4.6: Leaving out Ghungroo and Yorkshire, pairwise distances for various feature combinations reproduced.

BREED and feature	D-H
C: Colour	1.71
T: Texture	**1.83**
$T \cup C$	1.74
Best features	T
MAX-distances	1.83

node on the left happens to be Ghungroo (highest score) and the decision space is defined with respect the colour feature alone. Thus, the first level simplification results in a subtree (one versus three split), shown in Fig. 4.5. When the cumulative distances are recomputed (minus Ghungroo), one arrives at Table. 4.5 and from the deviation statistics, there is an indication that the tree structure is of type T3 (Fig. 4.4(c)) with the leaf being Yorkshire and the feature type being UNION of color and texture. Finally the tree structure converges to type T3 and the

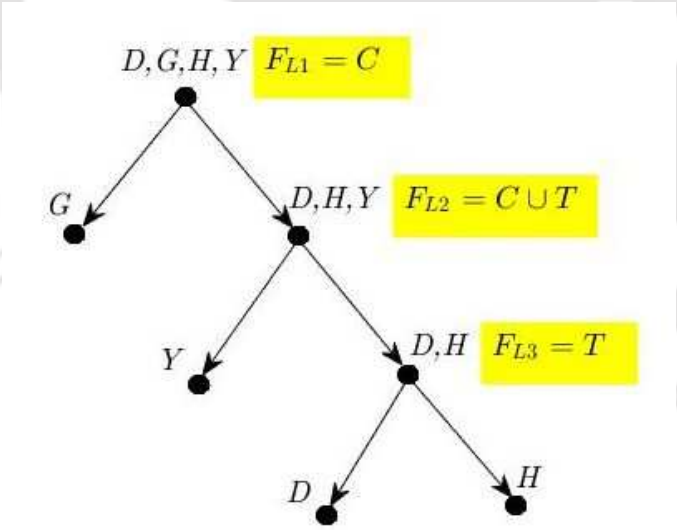


Figure 4.5: Final decision tree and siphoning policy.

final tree is shown in Fig. 4.5. Note in the last leg since it is Duroc vs Hampshire, the feature is Texture (T), Table. 4.6, as the highest score is registered for T). As per the siphoning order in Fig. 4.5, Ghungroo is separated from the rest with respect to colour, Yorkshire versus the rest with respect to colour/texture and finally Hampshire is separated from Yorkshire with respect to texture alone.

Table 4.7: Table showing the mean percentage accuracy for 100 iterations for the proposed colour, texture and combined features. A Flat Tree structure was deployed (i.e. linear classifier for a 4-class SVM).

	Colour		Texture		Combined	
	Mean	**St.Dev**	**Mean**	**St.Dev**	**Mean**	**St.Dev**
Duroc	68.14	21.80	51.88	18.00	84.00	13.85
Ghungroo	89.80	10.43	65.97	15.86	92.38	9.42
Hampshire	73.81	13.94	84.40	10.32	81.62	9.98
Yorkshire	95.44	6.13	96.49	2.97	98.22	3.57

4.3 Experimental results and comparisons

4.3.1 Database and training/testing procedure

The dataset used for the experiments consists of a total of 673 images for 55 animals taken from all the 4 breeds. Out of these 55 animals, 11 belonged to Duroc, 13 belonged to Ghungroo, 12 to Hampshire and 19 to Yorkshire. Thus on an average there were 12 images for each animal and these different images of the same pig shows the variation caused due to camera pan, skew effects and blur. 50% of the data was used for training and the remaining 50% for testing. This is a form of cross-testing, as the pigs used for testing are completely different from the pigs used for learning and generating the tree-classifier model (coupled with the optimal choice of features). For image acquisition, a high resolution digital camera was used. The images were cropped manually to suppress extreme background interference. But despite this we had to put these images through a level of automated segmentation, via a circular mask BALL-generation procedure to highlight portions only on the muzzle surface.

4.3.2 Performance evaluation for the Flat Tree

The formation of a decision policy to establish the optimal hierarchical scheme for classification is of paramount importance. The results for a 4-class linear classifier, are tabulated in Table 4.7. The mean and the standard deviation of the classification accuracies for 100 iterations are given for each of the three feature types.

Table. 4.7, shows that Yorkshire and Ghungroo have registered the highest numbers 98.22%

and 92.38%, respectively with respect to the composite (both colour and texture features), since, they are both distinct in terms of colour (Yorkshire all pinkish-white muzzle and Ghungroo is all greyish-black). This can be corroborated by the corresponding "Colour only" feature scores of 95.44% and 89.80% respectively (same Table. 4.7). The drop takes place with respect to "Texture only", for Ghungroo as it shares similar patch density profiles with Duroc and Hampshire (Ghungroo's classification accuracy drops to 65.97% for texture alone, while for Yorkshire it remains high at 96.49%). Since the pigs used for training and testing were completely different, the testing process was tough as the pigs exhibited considerable variability on the following fronts:

(i) Hair and pore density profiles: type, location and concentration of patterns was completely different for different pigs within the same breed; (ii) Breeds like Duroc and Hampshire which were expected to have pinkish-white patches on their muzzle, exhibited considerable unpredictability in the size, position and structure of the patch on the muzzle surface for the pigs being tested for the first time; (iii) These pigs being tested also exhibited differential variability with respect to local illumination and environmental settings.

4.3.3 Performance evaluation: Proposed TREE structure

The training and testing image subsets were randomly selected and the process was iterated 10 or more times to investigate the impact of training data variability on the final classification accuracies. At the beginning of each iteration, the training process involved computation of the features for those images and determining the optimal hierarchical structure (based on the algorithm in Section. 4.2.2) from the cluster/breed distances. Using the labeled features from the same training-arrangement, SVM classifiers were learnt for all the three stages of classification, identified by the feature type selection procedure. This random train-test splitting process was repeated multiple times. The hierarchy/TREE, obtained from the training data in all the 10 iterations, wherein the training and test sets were split randomly (to check tree stability), turned out nearly the same as shown in Table. 4.8, which indicated two things: (i) Tree structure was relatively insensitive to changes in the cross-validation arrangements during

Table 4.8: Hierarchy obtained along with accuracies using the proposed algorithm for the same 10 random selections of training data as used in Table 4.11 along with the feature choice at each node.

Iteration	**Linked List**	**Max-L1/** F_{L1}	**Max-L2/** F_{L2}	**Max-L3/** F_{L3}	**Duroc Acc.(%)**	**Ghungroo Acc.(%)**	**Hampshire Acc.(%)**	**Yorkshire Acc.(%)**
TREE-1	G-Y-(D,H)	5.82/C	3.81/(C,T)	1.8/T	91.12	93.52	86.71	96.23
TREE-2	G-Y-(D,H)	5.81/C	3.78/(C,T)	1.83/T	84.31	90.29	84.23	98.17
TREE-3	G-Y-(D,H)	5.77/C	3.8/(C,T)	1.77/T	76.23	87.32	78.55	96.78
TREE-4	G-Y-(D,H)	5.81/C	3.81/(C,T)	1.83/T	81.48	91.63	87.21	99.54
TREE-5	G-Y-(D,H)	5.83/C	3.77/(C,T)	1.79/T	88.45	94.79	80.48	100.00
TREE-6	Y-G-(D,H)	5.79/(C,T)	3.74/C	1.81/T	79.47	96.53	83.65	95.41
TREE-7	G-Y-(D,H)	5.8/C	3.79/(C,T)	1.82/T	80.45	89.41	90.32	99.54
TREE-8	G-Y-(D,H)	5.85/C	3.8/(C,T)	1.78/T	83.78	97.13	81.68	100.00
TREE-9	G-Y-(D,H)	5.77/C	3.82/(C,T)	1.82/T	77.21	86.33	92.57	100.00
TREE-10	Y-H-(D,G)	5.82/(C,T)	3.64/C	1.88/C	77.42	89.79	84.22	100

the train-test iterations, wherein 50% of the pigs were randomly picked for training and the other 50% were shortlisted for testing. Table. 4.8, shows the hierarchies obtained using the proposed algorithm for all the ten iterations. For seven out of ten (7/10) cases, the tree had the same structure: $G - Y - (D, H)$ (i.e. the breeds as leaf nodes were picked out in the order: Ghungroo, Yorkshire, and then finally Hampshire versus Duroc). (ii) The secondary features covering both texture and colour were robust for trapping breed-specific traits while dissolving individual variability within the same breed.

Table 4.9, shows classification accuracies with respect to the stable hierarchical scheme described in Section. 4.2.2 and shown in Fig. 4.5. In relation to the flat-tree (four-class) arrangement, the proposed TREE registered a higher score of 86.45% versus 84.00% for Duroc; 93.02% versus 92.38% for Ghungroo; 86.91% versus 81.62% for Hampshire and 98.54% versus 98.22% for Yorkshire; The main improvement was in the score for Hampshire (one of the toughest breeds for classification which exhibited a similarity in texture profile with respect to both Duroc and Ghungroo and also a similarity in colour with respect to Duroc). The improvement for Hampshire, stemmed from the fact that Ghungroo taken out earlier in the TREE arrangement, comparison could now be done exclusively on the texture front (ignoring the color profile).

Table 4.9: Classification accuracy (100 iterations), following the TREE hierarchy obtained, using our method.

	Mean	St.Dev
Duroc	86.45	12.05
Ghungroo	93.02	7.46
Hampshire	86.91	12.51
Yorkshire	98.54	4.84

The corresponding confusion matrix is shown in Table 4.10. The first row of the confusion matrix tabulates the number of actual Duroc muzzle images being classified as one of the four breeds. The second row gives the same numbers for the actual Ghungroo muzzle images. In the same way the third row corresponds to Hampshire and the fourth row to Yorkshire. Ideally the confusion matrix should have been a diagonal matrix with all the non-diagonal elements equal to zero. However, because of some mis-classification between the breeds, the matrix is not a diagonal one. From the first row of this confusion matrix, it can be observed that Duroc has the maximum confusion with Ghungroo on account of its moderate to high density of hair follicles and pores on the muzzle surface. This same reason applies to some of Ghungroo muzzle images being confused with Duroc muzzle images as observed from the first entry in second row. Also, it can be observed from Table 4.10 that there is some confusion between Duroc and Hampshire as well. This is because there is some similarity in the density profile of hair follicles and pores present on the muzzle surface of both the breeds. Also some of the Duroc breeds have a significant pinkish white patch on their muzzle which is mainly a characteristic of Hampshire breed and this becomes a source of confusion between the two breeds.

Table 4.10: Confusion matrix for proposed tree algorithm.

Predicted / Actual	Duroc	Ghungroo	Hampshire	Yorkshire
Duroc	56	5	3	1
Ghungroo	4	70	1	0
Hampshire	6	2	60	1
Yorkshire	0	1	1	108

4.3.4 Comparison with the Phylogenetic Tree algorithm (AGNES)

Table. 4.11, shows the hierarchies obtained from the Phylogenetic tree algorithm [44] used by the Phylogenetic MATLAB Toolbox [25] for 10 different random splits of training-testing data (50%-50%). The training-testing data used in the 10 iterations are the same as those used for evaluating the results in Table 4.8. While the Phylogenetic algorithm processes the breed-pair distance Table. 4.2, it is *feature agnostic*. It does not provide a mechanism for arriving at or selecting the optimal combination of features for forming the decision space at every intermediate node. Hence, for comparison purposes each of the node in the linked list has been supplied with the feature set $(C \cup T)$. Fig. 4.6, shows a typical tree produced by the AGNES MATLAB toolbox corresponding to the first row of Table. 4.11. The Phylogenetic tree structure (unlike the proposed algorithm) exhibits a high variability or data-sensitivity (four instances of TREE: $Y-G-(D,H)$; two instances of $G-Y-(D,H)$; two instances of $Y-H-(D,G)$ and two instances of $Y-D-(G,H)$) (Table. 4.11). In contrast, the proposed algorithm showed a strong base tree $G-Y-(D,H)$, which remained virtually the same (8 out of 10 times) for the same database-splits, identical to the proposed optimal tree in Fig. 4.8. Mean accuracies for AGNES toolbox are in Table. 4.12 based on the tree structure $Y-G-(D,H)$ (Yorkshire first and then Ghungroo, the main difference; last stage same, for both proposed and AGNES). Accuracies for Duroc, Hampshire and Yorkshire were on the lower side for AGNES, registering 83.58%,81.78% and 96.19% respectively. In contrast, the proposed TREE registered higher scores, 86.45%,and 86.91% for the critical/difficult breeds, Duroc and Hampshire respectively.

4.3.5 Comparison with Decision trees

Decision trees are useful when one is operating on raw measurements, as opposed to hand-crafted statistics and work via random attribute sampling, to identify the decision space where the separation between sub-groups of data is maximized. While the tree search is extensive, there are several reasons why this route turns out to be sub-optimal: (i) When features are handcrafted, they cannot be selectively dropped on a random basis, as every single statistic plays a crucial role towards breed separation. Deleting some of them from the decision making

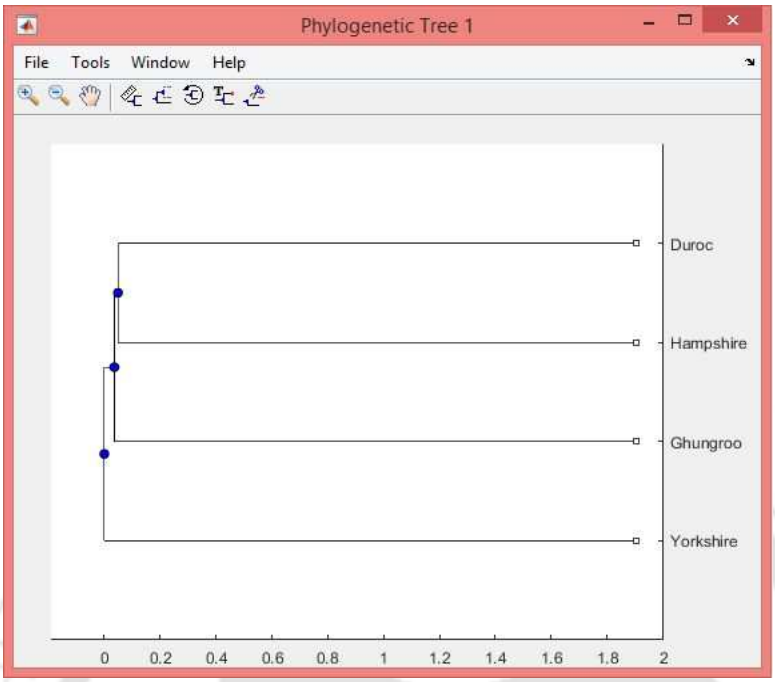


Figure 4.6: AGNES TREE-1.

Table 4.11: Table showing the hierarchies obtained from the AGNES algorithm used by the Phylogenetic Toolbox for the same 10 different random splits of training-testing data (50%-50%) as used in Table 4.8. Feature used at each node of hierarchy is $C \cup T$.

Iteration	Linked List	Duroc Accuracy(%)	Ghungroo Accuracy(%)	Hampshire Accuracy(%)	Yorkshire Accuracy(%)
TREE-1	Y-G-(D,H)	76.15	98.58	80.53	97.88
TREE-2	Y-D-(G,H)	83.21	90.22	75.48	97.34
TREE-3	G-Y-(D,H)	82.63	87.42	78.39	98.32
TREE-4	Y-G-(D,H)	80.51	89.78	77.24	100.00
TREE-5	G-Y-(D,H)	79.42	93.21	82.38	100
TREE-6	Y-H-(D,G)	85.43	92.19	84.71	97.23
TREE-7	Y-D-(G,H)	74.44	92.64	83.19	98.56
TREE-8	Y-G-(D,H)	77.83	90.19	76.22	100.00
TREE-9	Y-H-(D,G)	74.31	94.39	83.78	100.00
TREE-10	Y-G-(D,H)	86.48	100.00	82.64	99.14

Table 4.12: Accuracies obtained using the feature agnostic, Phylogenetic toolbox (features selected using the procedure from sub-section 4.2.2) and the mean tree $Y - G - (D, H)$ (Table. 4.11)

	Mean	St.Dev
Duroc	83.58	15.11
Ghungroo	94.11	9.07
Hampshire	81.78	11.92
Yorkshire	96.19	3.28

procedure or de-registering the feature vectors through random sampling, will only degrade the overall performance; (ii) Attributes when mixed as a union of colour and texture assume a "colourless" flavor, wherein no attribute is pre-labeled. No knowledge gained from the colour or texture feature analysis is used in the attribute search and sub-sampling procedure. In compact feature vectors, when tailor-made statistics such as the Sarle's bi-modality index [22] and the chromatic eigen value ratio, are randomly dropped from the (colour, texture) feature mixture, performance degradation is expected (as can be seen in Table. 4.13). Not surprisingly, the accuracies for Duroc, Ghungroo and Hampshire have dropped below 80%.

Table 4.13: Breed accuracies obtained using the decision tree architecture.

	Mean	**St.Dev**
Duroc	73.81	20.50
Ghungroo	75.78	17.76
Hampshire	71.42	14.84
Yorkshire	92.07	6.87

4.3.6 Absence of segmentation and Overall Picture

The proposed BALL-based muzzle segmentation process, described in Chapter. 2, ensures that colour and texture features are not affected by the portion outside the muzzle region. Comparisons between the proposed and classification algorithms from literature, in terms of accuracies, are now re-evaluated both with and without the mask. Table. 4.14 shows a reduction in the accuracies for the tree-based algorithms (including both proposed and the ones from literature), in the absence of segmentation. Note that interestingly the background profiles for different breeds are heavily pig-specific, as the background interference usually arises from two sides: (i) Face of the pig and (ii) In some cases, glove of the individual clutching the pig's snout. The fractional background portion outside the muzzle region constitutes roughly 25% of the total cropped square area. This background composition (25%) varies from pig to pig (even within the same breed), and creates an inconsistency in the breed-linked feature modeling procedure.

The impact of our BALL-based segmentation algorithm on the classification accuracy, is also checked and compared with the accuracies obtained by applying masks using the Chan-

Table 4.14: Accuracies for Proposed Tree, AGNES and Decision trees (NO segmentation).

	Proposed Tree		Phylo Tree		Decision Tree	
	Mean	**St.Dev**	**Mean**	**St.Dev**	**Mean**	**St.Dev**
Duroc	64.73	17.07	62.23	14.48	52.97	17.23
Ghungroo	81.78	13.39	79.50	17.76	69.38	17.89
Hampshire	80.77	11.23	80.96	8.02	66.71	18.30
Yorkshire	96.78	3.77	96.13	4.87	91.66	10.14

Table 4.15: Accuracies, with segmentation, using a variety of methods and comparison with proposed BALL-based approach (Difficulty level maximum for Hampshire breed).

	Proposed BALL-method		DRLSE		SDREL		Chan-Vese		NO segmentation	
	Mean	**St.Dev**	**Mean**	**St.Dev**	**Mean**	**St.Dev**	**Mean**	**St.Dev**	**Mean**	**St.Dev**
Duroc	86.45	12.05	80.29	9.72	66.33	18.37	51.72	19.34	64.73	17.07
Ghungroo	93.02	7.46	85.11	17.17	82.77	11.17	91.05	6.72	81.78	13.39
Hampshire	86.91	12.51	81.72	10.18	59.13	16.40	40.63	14.36	80.77	11.23
Yorkshire	98.54	4.84	93.68	4.84	95.11	6.39	87.10	9.30	96.78	3.77

Vese [30], DRLSE [29], SDREL [34], active contour models (in Table 4.15). Clear evidence of the superiority of the proposed BALL based segmentation conjecture, can be seen from the classification results, particularly for Hampshire, where the muzzle profile exhibits color diversity in a dual mode (i.e. pink and grey). Although the proposed segmentation algorithm is customized, it can still be used in segmentation tasks where the location of the contour defining the boundary of the object is approximately known; but is complex and partially diffused.

This work provides fairly strong evidence regarding the utility of the muzzle image of a pig as a potential pig breed identifier and can be extended further in the future with more number of breeds to further strengthen this hypothesis. The proposed tree synthesis procedure searches for a bias with respect to a specific breed (some breeds such as Ghungroo and Yorkshire are easier to pick out of the mixture) and chalks out, not just the selection order, but also the best choice of feature, at each intermediate decision point. Hence, this tree synthesis process, with many more secondary statistics and guidelines, can be extended to attack a larger frame

comprising of n classes, m basic feature types (m small), purely based on the pairwise distance table which will now have $O(2^m \times n^2)$ entries. The main weakness is that the current set of guidelines for tree-synthesis (based on cumulative, feature specific distance statistics) are fairly simplistic and effective for both small n (classes/breeds) and small m (basic feature types). For a larger arrangement, say large scale dog breed classification problem, based on facial images (large n and moderate m), one will require more sophisticated protocols and guidelines for intermediate decision making and feature selection.

5

Classification Scheme based on Outlier Detection Framework Using Maximum Spanning Trees

Contents

In the earlier chapter, a solution was developed based on the observation that when considering the problem of siphoning out individual breeds, it is important to decide the decision space in which this can be done. We showed that this selection rule can be formulated based on pairwise cluster distance values across different feature combinations and eventually leading to breed-specific cumulative distance statistics(i.e. i^{th} breed versus the rest $\forall i \in \{1, 2, 3, 4\}$). These cumulative distance statistics can then be used to build a tree which contains information regarding the order in which the breeds can be siphoned out and the corresponding feature type/combination deployed to form the decision space related to the SVM classifier at that node. Classification accuracies of around 86% obtained for both Duroc and Hampshire using the hierarchical classification scheme described in Chapter 4 provide scope for further improvement. It has been observed that the principal factors limiting the classification accuracies for the breeds Duroc and Hampshire are inter-class similarity and intra-class variation. This is shown in Fig. 5.1.

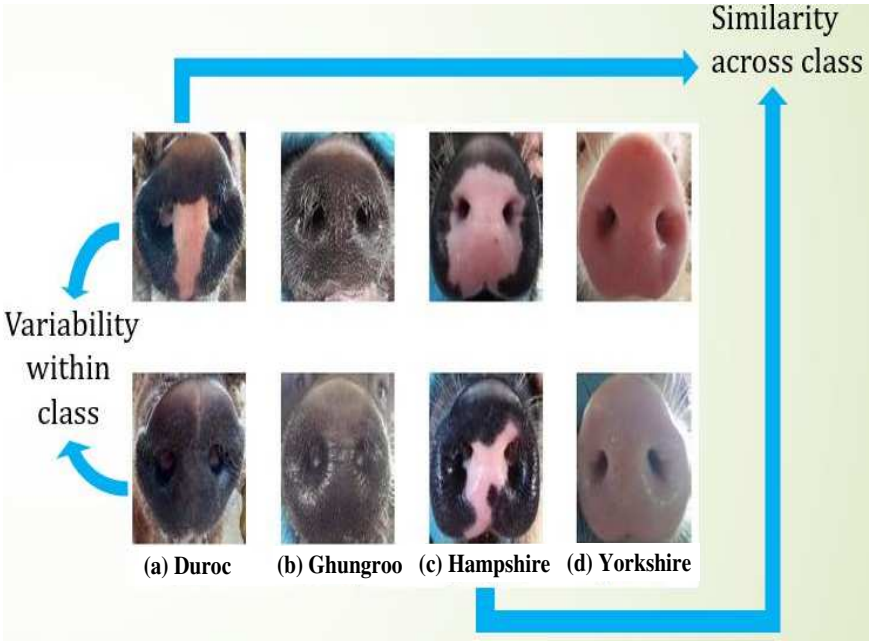


Figure 5.1: Exemplar muzzle images from different breeds showing inter-class similarity and intra-class variation.

Because of inter-class similarity, there is considerable overlap between clusters of feature

vectors from different classes, whereas intra-class variation leads to scattering of feature vectors from the training data of a particular class in feature space. When encountering such problems, it becomes necessary to assign likelihood or membership scores to a test feature vector (which falls in the overlapping region) corresponding to each of the overlapping training classes. There are two challenges here:

- When the training data is sparse and heterogeneous, it is important to stitch the data samples which are closely connected mainly to establish some form of reliability in the breed-specific training data representation. This can be done only via some form of spanning tree which is a data-centric macro-construct and binds the data together.
- While inclusiveness of a test feature vector can be established via a likelihood score the moment the tree-based data-centric model is available, it is important to weigh this likelihood against the likelihood scores generated with respect to tree models from other breeds.

When the feature vectors from a particular training class are distributed in feature space forming multiple compact clusters, it is advantageous to first identify each of these clusters and then learn suitable generative model for each of these clusters. Based on these, a classification scheme has been developed which uses an outlier detection framework for improving the classification accuracies in our breed classification problem.

When dealing with classification problems with limited datasets, a classic example of which is the dataset for our breed classification problem, it becomes necessary to bind the data samples from a given training class via some form of an association map. This motivated the search for a class of algorithms connected with spanning trees.

5.1 Multi-class classifiers and their limitations

When the number of data samples per class is very large and the representation is almost complete, the cluster in a higher dimensional space almost assumes an isometric structure. Thus every class-cluster can be modelled as an ellipsoidal blob of points [50]. The decision

boundaries between clusters are then hyperplanes [50], as one would find in a linear SVM model. Furthermore, if one attempts to generate multidimensional conditional probability density functions (PDFs) out of these class-samples, for a substantial class representation, these conditional PDFs will assume a Gaussian form [51] with their own mean vectors and class-covariance matrices. This lends itself to decision rules involving log-likelihood ratio tests etc. [52] which can be quite effective provided the class cluster density profile exhibits some form of contiguity and gradual variability about the mean vector. However, when the data-sample representation is not complete and the volume of data per class is moderate, the clusters will not be contiguous in space. Under these circumstances class detection becomes purely a geometric exercise, as opposed to a conditional likelihood ratio test.

As mentioned earlier, the current work connected with pig breed analysis involves four breed-classes viz. Duroc, Ghungroo, Hampshire and Yorkshire [53], amongst which Ghungroo is a breed found commonly in the North-Eastern part of India [54], while the others are imported breeds in India [55]. Along with the intra-class variability and inter-class similarity already mentioned, the classification accuracy for breeds such as Duroc and Hampshire is limited because of the relatively small size of the dataset available for breed classification. On an average, each class (among the four distinct breed-classes over which data was gathered) had anywhere between 10 and 15 pigs (with 10-variations per pig). Out of these 10-15 pigs per class, only 50% could be used for training which imposed severe constraints on the training procedure. This limited training data ruled out the usage of Neural Networks (Multilayer Perceptrons) and Deep Neural Network architectures [56]. These architectures are commonly used when labelled training data are available in large numbers as in [57] or auto-population can be done with the limited training data [58]. However, when the intra-class variability is high (i.e. breed specific data exhibit a high degree of variability with respect to colour and texture) and the training samples available per breed is as low as 50-70 (with 5-7 pig-subjects per breed), auto-population both at the image level as well as the feature level is ruled out. SVMs [59] could not be used because of the overlapping nature of data from different classes(particularly for Duroc and Hampshire). In the case of overlapping clusters there is no conflict resolution

protocol or a soft-metric which assigns likelihood scores(to feature points) corresponding to each cluster. It is precisely why an association map between feature-points constructed based on feature similarity becomes useful to form a much more clear opinion regarding the samples in the intersection region [60]. In the case of Nearest Neighbour algorithms [61], since the voting is purely based on proximity and less on the overall relative structural arrangement of the training data-points, the bigger picture is lost and it remains sensitive to noise.

When confronting large intra-class variability and limited training data, the first step was to stitch together the training samples via some form of association map or feature proximity map so that one could gain a better understanding of the spread of this data particularly the span, shape and directional growth. This motivated the search for a class of algorithms connected with spanning trees. In [62, 63] Minimum spanning trees (MinST) have been deployed in an unsupervised mode to detect the outliers from a regular set of measurements. In [64–66] Minimum Spanning Trees(MinST) have also been used for outlier detection when built over limited datasets. The main issue with this body of work was the choice of soft-statistic related to a particular test point. The degree to which a test point belonged to a training cluster was decided on the basis of the proximity of that point to the closest edge connecting nodes in the MinST. There were some issues with this frame:

- Since outlier scoring was primarily based on local neighbourhood information about the test-point, the bigger picture is lost. Questions such as: Is the test-point contained within the convex-hull of the training cluster and if so, how deeply embedded is it?, remain unanswered.

- In a multi-class setting there could be overlapping classes, which in turn implies that a test-point could have multiple owners. To measure the degree and extent of ownership, one needs a more reliable scoring mechanism to resolve this INLIER-INLIER conflict.

The current work revolves around the development of a multi-class classifier using an outlier detection scheme based on Maximum Spanning Trees(MaxST) [67].

There are some applications in literature which use the MaxST formulation, particularly

because connectivity based on diversity (dissimilarity) is emphasized. In [68], the authors used MaxST to construct a Media Distribution Tree for efficient streaming of media between mobile smart devices. In [69], a set of attributes are selected from a given feature set to eliminate the redundant attributes. The redundant attributes are the ones which have a high correlation with other attributes in the feature set. In the same light, MaxSTs also find applications in the compression of MODIS [Moderate Resolution Imaging Spectroradiometer] (which is an instrument aboard polar satellites that provides earth observing data in different spectral bands) data [70]. While analyzing the nodal inter-connectivity in brain networks, MaxSTs present a robust and compact description of this inter-connectivity profile [71]. It was observed in Wang et al. [71] that spurious interconnections which are likely to surface or re-surface over a course of time in a subject, are less likely to influence the base-tree.

In the context of pig-breed analysis, for every breed-class the dominant sub-clusters are identified and then robust MaxST tree-models are learnt for each of these breed-specific sub-clusters. Similar collective MaxST models are built for all the four pig breeds (Duroc, Ghungroo, Hampshire and Yorkshire). When a test point is confronted, an outlier score is generated corresponding to each of the individual classes based on the change induced by the test feature point to on the MaxST representation. The test feature point is assigned the class label for which it has the lowest probability of being an outlier.

Each class present in the training data is treated as a target class and a spanning tree representation is learnt for the same. In the following section, the spanning tree representation learnt from the training data corresponding to each class is discussed.

5.2 Tree Generation and Graphical Model Building for Outlier detection

The training process in the proposed outlier detection scheme can be broadly divided into two parts: The first part involves identifying multiple *compact* clusters in the spatial distribution of training feature points/nodes. In all subsequent discussions the terms '*feature point*' and '*node*' will be used interchangeably as they refer to the same thing in the context of this

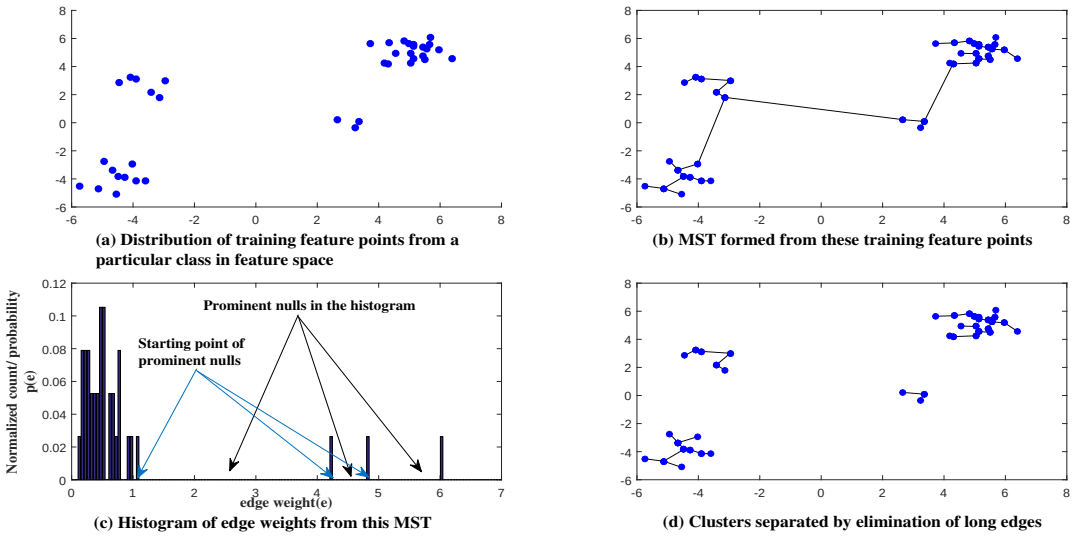


Figure 5.2: The process of identifying the clusters present in the training samples of a particular class.

discussion. The second part is concerned with the learning of MaxST representation for each of these *compact* clusters.

5.2.1 Forming Compact Clusters from training data

Due to diversity within the target class, it may happen that the feature points exist in multiple *compact* clusters. A test feature point which does not lie within or in the vicinity of any of these clusters is a probable outlier with respect to the target class. Thus it becomes important to break the target class into its constituent clusters such that each cluster is *compact* in itself. Assuming there are k nodes in a cluster, let d_{NN_i} denote the nearest neighbour distance of the i^{th} node in the cluster. If the cluster is *compact*, then it is expected to satisfy

$$1 < \frac{\underset{i}{\text{MAX}} d_{NN_i}}{\underset{i}{\text{MIN}} d_{NN_i}} < 1 + \epsilon \tag{5.1}$$

where, $\epsilon > 0$ is a small positive real number. The MinST has been used extensively in literature [63, 72–74] to identify the clusters in the training data due to its simplicity of implementation and efficacy, purely based on a distance profile analysis. Inspired by this, as a pre-processing and data-massaging step we have used MinSTs to detect the key clusters present

in a particular breed class. Fig. 5.2(a) shows a set of training feature points from an arbitrary target class. Fig. 5.2(b) shows the MinST constructed from these feature points. In this MinST representation, the weight of an edge connecting two feature points is the Euclidean distance between them. It can be observed that there exist more than one compact cluster in this training set, and any two such clusters are joined to each other by a single long edge in this MinST representation while the feature points within a cluster are joined by short edges. Thus all the edges in this MinST can be categorized into two classes: viz. longer edges connecting feature points across clusters and shorter edges connecting feature points within a cluster. This binary categorization of the edge weights is precisely the reason why an MinST representation has been chosen for the purpose of separating the clusters. The longer edges connecting feature points across clusters need to be eliminated to identify the clusters. Also, if there are n_c clusters then they are connected to one another by exactly $n_c - 1$ edges. Fig. 5.2(c) shows the histogram of all these edge weights. Since the edge weights connecting feature points across clusters are large and distinct, hence we get *prominent nulls*(i.e. large sequence of zero weighted bins between two non-zero weighted bins) in the histogram as shown in Fig. 5.2(c). These nulls become more prominent as the inter-cluster distance increases with respect to the largest distance between any two feature points within a cluster. Hence, a suitable threshold(e_T) is computed which is a function of the position of these nulls. Based on this threshold value, the relatively long edges between two different clusters can be removed to obtain n_c connected components(corresponding to n_c clusters) as shown in Fig. 5.2(d). The procedure adopted to determine this threshold is described next.

The histogram of the edge weights shown in Fig. 5.2(c) is examined to determine the starting position of each *prominent null*. A successive string of zero-weighted bins in the histogram is considered as a *prominent null* if this successive number of zero-weighted bins exceed a predefined threshold. Assuming m prominent nulls are present in the histogram (for our example $m = 3$), let $e_{z1}, e_{z2}....e_{zm}$ denote the edge weights corresponding to the starting point of each of these prominent nulls. Then $e_T = MIN\{e_{z1}, e_{z2}....e_{zm}\}$, such that the cumulative distribution

function CDF$\{e_T\}$ is greater than a threshold (usually in the range of 0.9 or above). The choice of threshold is closely connected to the priori estimate related to the number of distinct clusters [73]. This is to prevent the formation of spurious clusters and weed out noisy points midway between clusters.

5.2.2 Generating outlier detection scores based on Maximum Spanning Tree analysis

When a test sample is received and positioned with respect to the training feature set of a specific breed, which may comprise of n_c compact clusters, one of the following scenarios may arise:

S1 The test point could be outside all the n_c clusters (exterior case).

S2 The test point could be sitting inside one of the n_c clusters (interior case).

If positioning corresponds to the scenario S1, i.e. exterior case, then, the proximity of a test feature point to each of these n_c clusters need to be estimated first. To get a measure of this proximity, separate MaxSTs are constructed from the fully connected graph of each of these n_c clusters. Any edge in the MaxST connecting two feature points is the Euclidean distance between them.

The following analysis corresponding to scenario S1, reveals as to how the proximity of a test feature point to one of the n_c compact clusters can be estimated via a MaxST analysis. A MaxST has been chosen for representing the distribution of training feature points within a cluster for the following reasons:

- For a given set of nodes forming a cluster, which are fully connected with distinct edge weights, the MaxST forms a unique representation.
- If the test point is inside the cluster, it has been observed that the MaxST constructed by inducting the test point with the remaining nodes, remains the same when viewed with reference to the training nodes. However if this test point is outside the cluster, there is a change in the tree structure with respect to the training nodes, which can be quantified.

- The extent of change caused in the structure of this MaxST can be shown to be a function of the proximity of the new test node to the cluster under scrutiny.

This is illustrated in Fig. 5.3 where a set of training feature points compactly arranged in feature space from the original cluster and a test feature point forms the new node. There are several established methods in literature for MinST construction from connected graphs. The classical Prim's algorithm [75] has been chosen for our purpose. A MaxST can be constructed using the Prim's algorithm by feeding the Prim's algorithm with the set of nodes and the negative of the edge weights connecting the nodes.

Prim's algorithm is a greedy algorithm and constructs the MinST in an iterative manner. The algorithm starts with an empty tree TR containing no nodes and edges. Two different sets are maintained during the course of the algorithm: N which is the set of nodes which are yet to be added to TR and W is the set of edges from the connected graph which have still not been added to TR at the end of an iteration. In the first iteration a node is selected at random from the set N and added to TR. Thus, TR becomes a single node tree. The second iteration starts with searching for an edge in W with minimum edge weight such that it connects the node in TR. This edge and the node from N which is connected to the other end of this edge are now a part of TR and hence are removed from W and N respectively. This process is iterated till all the nodes in N are connected to TR, thus making TR a MinST. At every iteration while choosing a new edge from W, it is imperative to ensure that no loops are formed in TR.

Table 5.1 lists the notations that will be used for the rest of the paper. The weights of the edges that exist between nodes in a MaxST can be put in the form of an adjacency matrix [76]. The element of an adjacency matrix A with index (i, j) i.e. $A(i, j)$ is zero, if there does not exist an edge between nodes i and j in the MaxST or $i = j$; else $A(i, j) = w_{i,j}, i, j \in \{1, 2, ...k\}$. The different adjacency matrices to be used henceforth are denoted by symbols listed in Table 5.1. The elements in the last row and column of $A_{S_{EXT}}$ contain the weights $w_{T,i}, i \in \{1, 2, ...k\}$. Thus, the sub-matrix $B_{S_{EXT}}$ contains information regarding edges in $MaxST_{EXT}$ which have been either preserved or removed involving only the training nodes in S.

With k nodes in the cluster, A_S and $B_{S_{EXT}}$ have dimensions of $k \times k$ each, while, $A_{S_{EXT}}$ has

Table 5.1: Notations used.

Symbol	Property of cluster denoted by symbol
S	Set of training nodes in a particular cluster.
k	Number of nodes in the cluster.
T	Test node.
i	Node with index i in S
$e(i,j)$	Edge connecting any two nodes i and j
$w_{i,j}$	Weight of edge connecting nodes i and j
$e(T,p)$	Edge connecting the test node T with a specific node p
$w_{T,p}$	Weight of edge connecting test node T with node p
$MaxST_{ORIG}$	MaxST formed by the set of nodes in S.
$MaxST_{EXT}$	MaxST formed by including T with the nodes in S
A_S	Adjacency matrix corresponding to $MaxST_{ORIG}$
$A_{S_{EXT}}$	Adjacency matrix corresponding to $MaxST_{EXT}$
$B_{S_{EXT}}$	Sub-matrix obtained from $A_{S_{EXT}}$ by deleting its last row and last column.
Σ_S	Sum of the edge weights in the tree $MaxST_{ORIG}$
$\Sigma_{S_{EXT}}$	Sum of the edge weights in $MaxST_{EXT}$
E	Set of edges in $MaxST_{ORIG}$

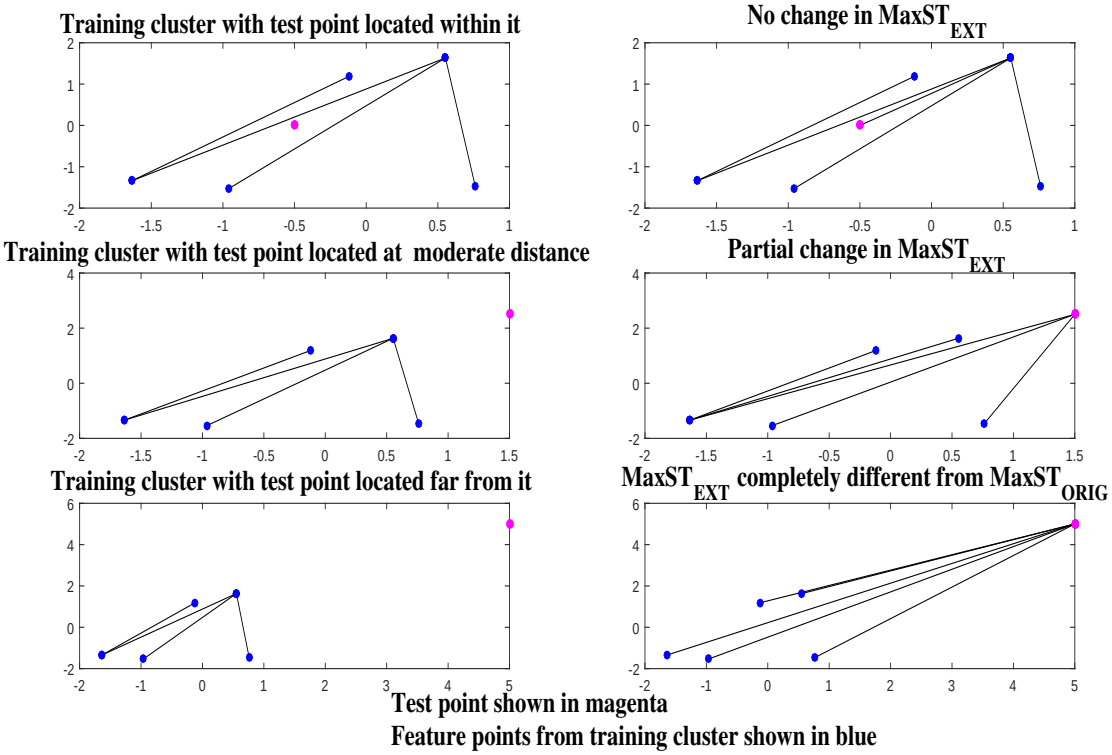


Figure 5.3: Effect of test feature point on the MaxST representation with respect to the nodes of the training cluster.

dimensions of $(k+1)\times(k+1)$. The k nodes in $MaxST_{ORIG}$ are connected by $k-1$ edges, while the $k+1$ nodes in $MaxST_{EXT}$ are connected by k edges. In the feature space, as the Euclidean distance of T with respect to the nodes in S becomes larger, its probability of becoming an outlier will increase. It will be shown that the change in the structure of $MaxST_{EXT}$ with respect to $MaxST_{ORIG}$ increases progressively, as a test feature point moves farther away from the nodes in S. A metric δ is defined to bring out the relative difference between the tree structures: $MaxST_{ORIG}$ and $MaxST_{EXT}$ as follows:

$$\delta(T) = \frac{||A_S - B_{S_{EXT}}||_F}{||A_S||_F} \tag{5.2}$$

The behaviour of $\delta(T)$ as the location of test node T changes in feature space needs to be investigated with $0 \leq \delta(T) \leq 1$. It is necessary to examine the conditions under which the outlier score assumes the two extreme values zero and one.

CLAIM::

- **CL-1**: For any test point in the interior of the training cluster or within the convex hull of the training set, $\delta = 0$.
- **CL-2**: The value of δ increases on an average monotonically from zero to one, as the distance of the test point from the cluster centroid increases. The distance saturates beyond a certain point d_{sat}, which a function of the cluster structure.

Proof: Firstly given any cluster of k points, this can be partitioned into two disjoint sets of points: (i) Points or nodes which form the boundary of the convex hull of the cluster (i.e. the points which are at the periphery, $S_{BOUNDARY} = \{P_1, P_2, ..., P_{k_1}\}$) and (ii) Points which are in the interior of the cluster, $S_{INTERIOR} = \{Q_1, Q_2, ..., Q_{k_2}\}$, such that,

$$\begin{aligned} S &= S_{BOUNDARY} \bigcup S_{INTERIOR} \\ S_{BOUNDARY} &\bigcap S_{INTERIOR} = \phi \\ k_1 + k_2 &= k \\ & P_i, Q_j \in S \end{aligned} \tag{5.3}$$

While constructing the MaxST, every point in the interior of the convex hull $Q_j \in S_{INTERIOR}$, will connect with one of the boundary points $P_i \in S_{BOUNDARY}$ only (irrespective of its position in the interior). It will not connect with another interior point. This subsequently has the following implications:

- If T is a test point within the convex hull of the training cluster and $MaxST_{ORIG}$ is the maximum spanning tree generated with respect to the k training nodes, the δ score generated with respect to the maximum spanning tree constructed over the extended set: $S_{EXT} = S \bigcup T$, i.e. $MaxST_{EXT}$ is zero. This is because since T is in the interior of the hull, no training node will connect with T. The connections of all the nodes in the training set will take place with respect to the boundary points alone $P_i \in S_{BOUNDARY}$. Thus the base graph structure will remain same and furthermore,

$$degree(T \text{ in } MaxST_{EXT}) = 1 \tag{5.4}$$

Thus, claim, CL-1 is proved and it follows that,

$$A_S(i,j) = B_{S_{EXT}}(i,j), \forall i,j \in \{1,2,3...k\} \tag{5.5}$$

wherein the first k nodes involved in both the incidence matrices are the pre-labelled training nodes. This same theoretical result has been verified empirically by computing the δ scores over a particular training set. This is illustrated in Fig. 5.4, wherein the training points are shown in blue and all the test points which register zero δ scores are in green, while the test points which register $\delta > 0$ scores are in red. The convex hull for this training cluster along with the boundary points are indicated via a dotted line. All the chosen test points within the convex hull produce zero δ scores, as seen from the green dots within the hull (Fig. 5.4).

Consider a set of nodes $i \in \{1,2,...k\}$ from S (shown in blue) and a test node T (shown in red) distributed in $\mathbb{R}^2$ feature space as shown in Fig. 5.5(a). The MaxST $MaxST_{ORIG}$ constructed from the nodes in S are also shown in Fig. 5.5(a). Fig. 5.5(b) shows the MaxST $MaxST_{EXT}$ formed by including T with the nodes in S. One edge in $MaxST_{EXT}$ (i.e. $e(1,4)$)) is not preserved, while two edges are formed in $MaxST_{EXT}$ connecting T with the nodes in S.

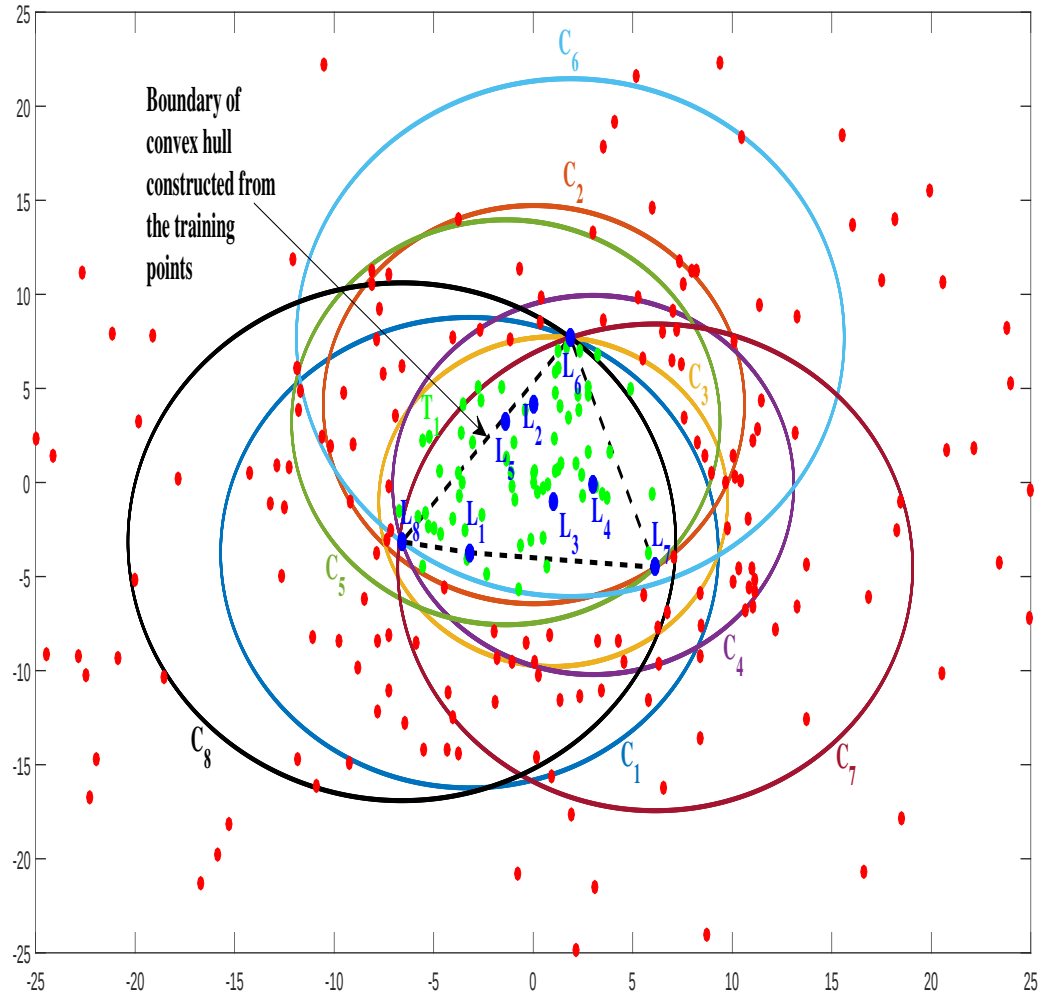


Figure 5.4: Region in feature space for which $\delta = 0$ shown as intersection of circular regions. Taking each node in set G as the centre a circle is drawn whose radius is the maximum edge weight of the edge connecting that node.

Test point positions for studying the impact of increasing distance and the angular positioning is shown in Fig. 5.6. This impact has been discussed via two plots: (i) Referential δ scores produced for different radii, when the test point is moved radially outward from the centroid of the cluster along the positive X-axis in increments of d_{min} (Fig. 5.7(a)), where,

$$d_{min} = MIN_{i,j \in S; i \neq j} D_S(i,j) \tag{5.6}$$

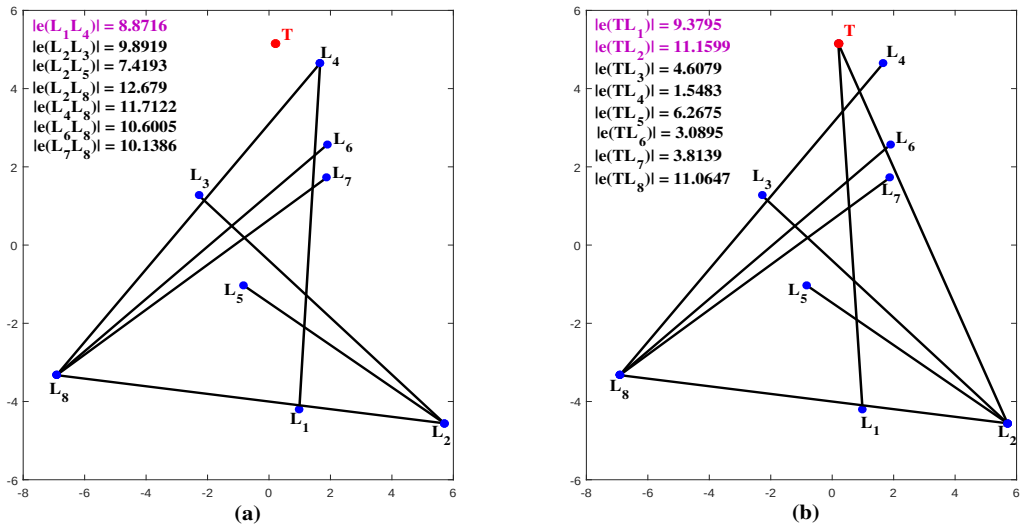


Figure 5.5: Changes taking place in the structure of the MaxST when the new test node T is inducted.

(ii) For the same radial increments, the delta·scores are computed for radial distances, $r_k = k \times d_{min}$ and random angular positions $\theta_k \in [0, 2\pi]$. The standard deviations are computed as function of the radial distance in Fig. 5.7(b). On a mean scale, it is easy to observe from Fig. 5.7(a), that the δ scores increase monotonically as the point moves out of the cluster and away from the cluster centroid. When the test point falls outside all the circular zones marked in Fig. 5.4, the score becomes exactly unity.

5.2.3 Region for which $\delta = 0$

For a particular training set, let j_T be the farthest neighbour of T in $MaxST_{EXT}$. In the formation of $MaxST_{EXT}$, an edge $e(i, j)$, with reference to node $i \in S$ is preserved in $MaxST_{EXT}$, when the edge weight $w_{T,i}$ is less than the maximum of the edge weights $w_{i,j}$ taken with reference to i and over all j with $e(i, j) \in MaxST_{ORIG}$. This is true for all the nodes i in S except j_T. Thus it can be inferred that $\delta = 0$ if

$$w(T, i) < w^i_{MAX} \forall i \in \{1, 2, ...k\}$$
$$w^i_{MAX} = \underset{j}{\text{MAX}} w_{i,j}$$
$$j \neq i, e(i, j) \in MaxST_{ORIG} \tag{5.7}$$

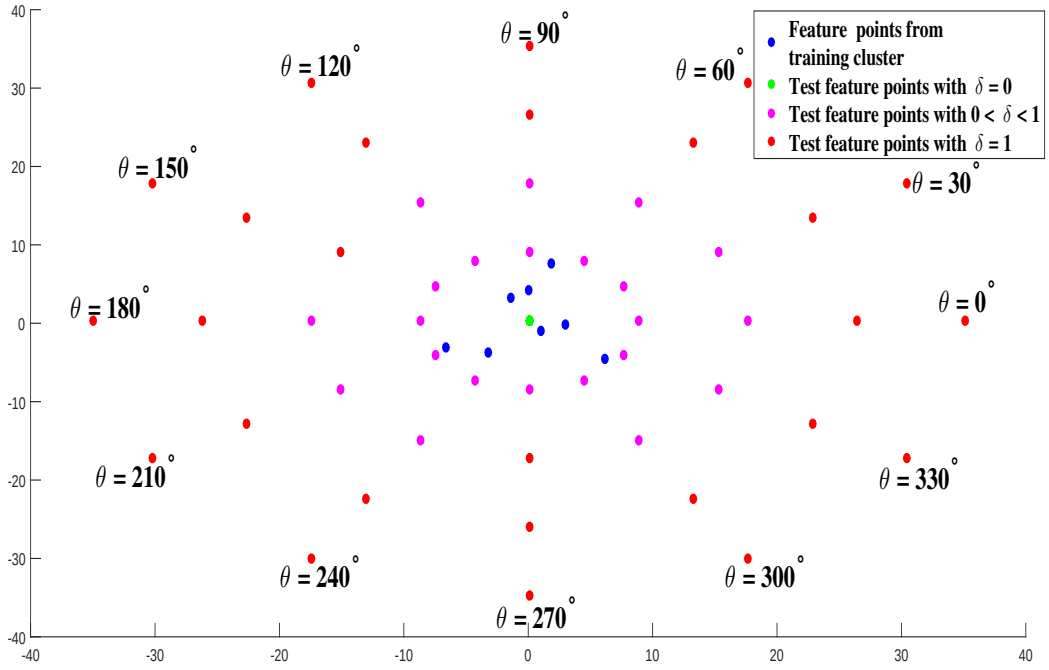


Figure 5.6: Test point positions for studying the impact of increasing distance and the angular positioning.

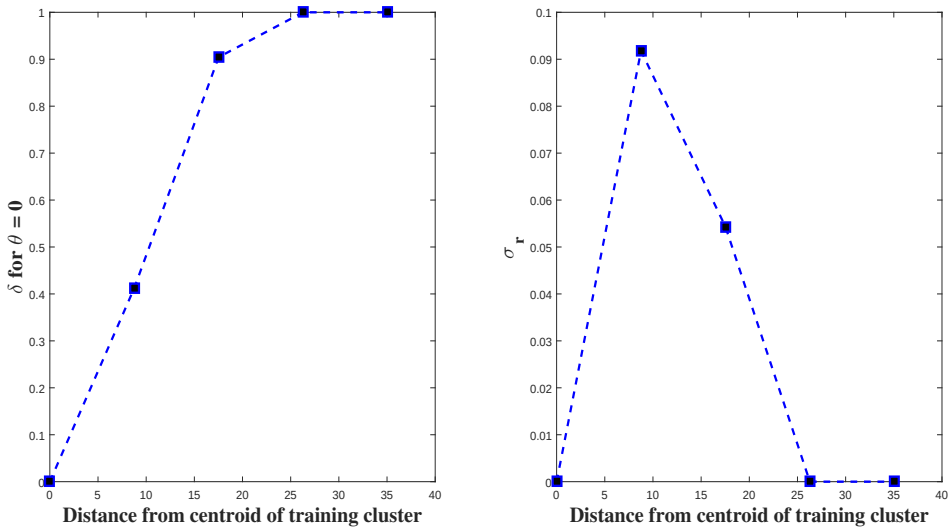


Figure 5.7: Impact of increasing radial distance of test point from the cluster centroid on the *delta*-scores.

i.e. the nodes in S do not connect with the test point T, while T connects with exactly one node, i.e. j_T. This subsequently implies that the degree of T remains '1'.

In the light of (5.7), a circle C_i is drawn taking the location of the node i as the centre and w^i_{MAX} as the radius. If similar circles are drawn for all the nodes in S, then given a test node lying outside the region of common intersection of all such circles (as shown in Fig. 5.4), some of the edges in $MaxST_{ORIG}$ should not be preserved in $MaxST_{EXT}$. That is δ should not be zero in such a case. This point is illustrated in Fig. 5.4, where the nodes shown in blue form the set S in $\mathbb{R}^2$ feature space and are labelled as $i \in \{1, 2, ...8\}$. A number of test nodes are distributed in this feature space. The test nodes for which δ evaluates to zero are shown in green; while the others are shown in red. It can be observed that δ evaluates to zero for all the test nodes lying in the common intersection region of the circles $C_i, i = 1, 2, ...8$. δ is found to be non-zero for most of the test nodes lying outside this intersection region.

The test node labelled T_1 is such an exception. This happens because for T_1, $i = 7$ happens to be the farthest node in S and T_1 also lies only outside the circle C_7(and not outside any other circle $C_i, i \neq 7$). Hence, T_1 cannot connect itself to any node other than $i = 7$ and replace any edge connected to the corresponding node, because that would destroy the maximum sum property of $MaxST_{EXT}$. Thus if a test node T lies in the interior of all circles except one, say C_i then $\delta = 0$ under the condition that the farthest neighbour of T is the node i.

It is also to be noted that if any test node T lies in the exterior of more than one circle, then δ always evaluates to a non-zero value. In such a situation, there always exists an edge in $MaxST_{ORIG}$ which can be replaced, thus making $\delta \neq 0$. When the location of T is such that it lies outside all the circles, then $B_{S_{EXT}} = \mathbf{0}$(null matrix) and hence $\delta = 1$ because none of the edges in $MaxST_{ORIG}$ is now preserved in $MaxST_{EXT}$. All the nodes in S connect themselves to T and this situation was shown earlier in Fig. 5.3(third row, second column).

Although this analysis was carried out in $\mathbb{R}^2$ space, it holds for higher dimensional spaces too. In $\mathbb{R}^n (n > 2)$ space, the circles will be replaced by hyper-spheres.

5.2.4 Score adaptation accounting for anomalous cases

Now consider the location of the test node T with respect to the nodes in S as shown in Fig 5.8. In this case, although the metric δ evaluated for T is zero, it is a highly probable outlier with respect to the nodes in G. Thus, the nearest neighbour distance of the test node T with respect to the nodes in G also plays a crucial role in determining whether T is an outlier with respect to the nodes in G. When the test node T is at a very large distance from the cluster G(third row in Fig. 5.3), δ saturates to a constant value of one and does not increase with further increase in the distance of the test node. In such a case also the nearest neighbour distance becomes an important metric to determine the probability of T being an outlier. Thus, in order to determine the probability of a node T being an outlier, the metric δ is modified as,

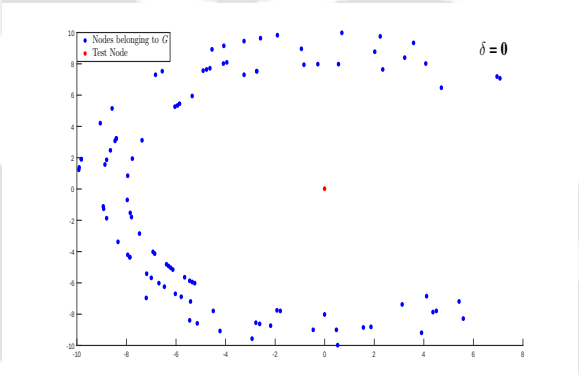


Figure 5.8: Test node being an outlier although it is a convex combination of the nodes in G.

$$\Delta = e^{\delta} \times d_{T,min} \tag{5.8}$$

where $d_{T,min}$ is the nearest neighbour distance of T with respect to the nodes in S.

5.2.5 Additional notes

With $MaxST_{ORIG}$ and $MaxST_{EXT}$ representing MaxSTs, the metric δ is zero when T is an inlier with respect to the nodes in S. As the location of T falls outside the cluster formed by the nodes in S, δ becomes non-zero and starts increasing as T moves farther away from this cluster till it saturates to a value of one. Such a systematic behaviour cannot be observed if MinSTs were deployed instead of MaxSTs.

5.3 Outlier Detection Framework applied to multi-class classification.

In a multi-class classification scenario, the test feature point is assigned the class label of that class, for which it has the lowest probability of being an outlier. The binary classification problem is considered first for the sake of simplicity. It is assumed that the feature points from both the training classes form a single compact cluster. As shown in Fig. 5.9(a), depending on whether the test feature point lies completely surrounded by the training feature points of a target class or not, it can be termed as either an inlier or outlier respectively with respect to that class. With this notion, for a binary classification problem, four possible cases arise, which are:

- *Case-1:* When the test feature point is an outlier for both the classes.
- *Case-2:* When the test feature point is an inlier for class C_1 but outlier for class C_2.
- *Case-3:* When the test feature point is an outlier for class C_1 but inlier for class C_2.
- *Case-4:* When the test feature point is an inlier for both the classes as shown in Fig. 5.9(b).

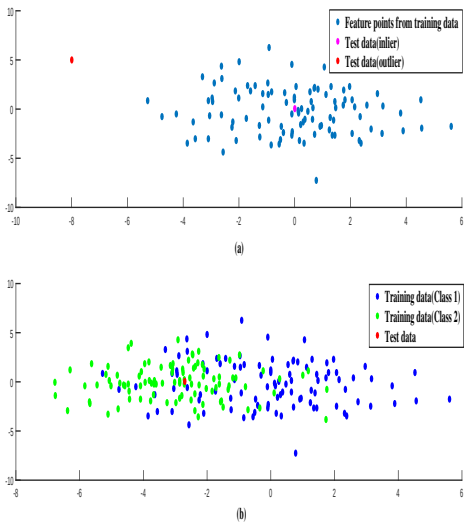


Figure 5.9: (a)Examples of inlier and outlier test feature points. (b)Instance when test feature point becomes an inlier for two overlapping classes.

Out of these four cases, *Case-4* requires special attention; and this situation arises when the test feature point lies in the overlapping region between the training feature points of two classes as shown in Fig. 5.9(b). In this case, the test feature point can be claimed as belonging to either of the two classes. This conflict regarding the membership of the test feature point to a particular class can be resolved by examining how deep the test feature point lies within the training feature points from each of the two classes. In such a situation, the test feature point is assigned the membership of that class corresponding to which it is more deeply embedded within the training feature points. At this point we digress to find out a metric which gives an estimate of how deeply a test feature point lies within a particular class.

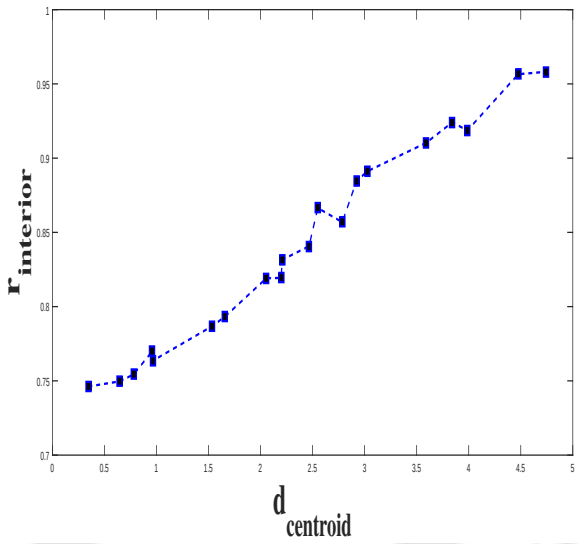


Figure 5.10: Variation of $r_{interior}$ with distance from centroid($d_{centroid}$) of training feature points.

Consider the scenario in which multiple test feature points are embedded within the distribution of training feature points from the target class shown in Fig. 5.9(a). For each of these test feature point $\delta = 0$ with respect to the target class. Let N_{TR} denote the number of triangles formed by choosing triplets of feature points from the training data. For a training set with k nodes, $N_{TR} = \binom{k}{3}$. Let $N_{interior}$ denote the number of triangles which have the test feature point in their interior. The deeper the test feature point lies within the training data, the higher the value of $N_{interior}$ will be. The methodology used to determine whether a test point lies inside a triangle whose vertices are given is described in [77]. The metric $N_{interior}$ is

normalized with respect to N_{TR} to make it invariant to the number of training feature points within a class. A normalized metric is defined as:

$$r_{interior} = 1 - \frac{N_{interior}}{N_{TR}} \tag{5.9}$$

A plot of $r_{interior}$ versus the distance of the test feature point from the centroid of the training cluster $d_{centroid}$ is shown in Fig. 5.10. As the test feature point moves farther away from the centroid, $N_{interior}$ keeps decreasing and hence $r_{interior}$ keeps increasing. Thus $r_{interior}$ gives an indication of how deep the test feature point lies within the training data. The metric $r_{interior}$ has been developed when the feature points are distributed in $\mathbb{R}^2$. For the more general case when the feature points are distributed in $\mathbb{R}^n$, $r_{interior}$ can be evaluated by projecting the feature points from $\mathbb{R}^n$ onto $\mathbb{R}^2$ space by using a random projection matrix [78]. Random Projection Matrices(RPM) have been preferred over other dimensional reduction methods because they can more accurately preserve the Euclidean distances between feature points in lower dimensional projected space as compared to other dimensional reduction methods like Principal Component Analysis or Discrete Cosine Transform [79]. While projecting the feature points from the higher dimensional space onto the lower dimensional space, the RPM introduces some noise in the distances between feature points in the projected space. This noise can be reduced by taking multiple projections of the feature points in the lower dimensional space using multiple RPMs and averaging the distances between feature points. Consider a set of training feature points forming a compact cluster in $\mathbb{R}^{18}$. The choice of $\mathbb{R}^{18}$ as the higher dimensional feature space is motivated by the fact that the feature points f^{MUZZLE} generated for the pig breed classification problem(as will be discussed later in Section 5.4) lies in $\mathbb{R}^{18}$. All the feature points in $\mathbb{R}^{18}$ are projected onto $\mathbb{R}^2$ using n_{in} different random projection matrices. Considering any pair of two feature points in $\mathbb{R}^{18}$, the ratio r^{ij} is defined as

$$r^{ij} = \frac{d^{ij}_{\mathbb{R}^{18}}}{d^{ij}_{\mathbb{R}^2}} \tag{5.10}$$

where $d^{ij}_{\mathbb{R}^{18}}$ and $d^{ij}_{\mathbb{R}^2}$ are the distances between feature points i and j in $\mathbb{R}^{18}$ and $\mathbb{R}^2$ feature

space respectively. This ratio needs to be constant $\forall i, j$, so that the cluster in $\mathbb{R}^{18}$ is properly projected in $\mathbb{R}^2$. However, because of noise introduced by the RPM, r^{ij} is not maintained constant $\forall i, j$ and there is some non-zero standard deviation associated with it. As Fig. 5.11 shows, the standard deviation in r^{ij} can be reduced by taking multiple projections of the feature points. After projecting the feature points in $\mathbb{R}^2$ for n_{in} different instances using n_{in} different RPMs, the distances between each pair of feature vectors is averaged over n_{in} instances. These mean distances between each pair of feature vectors r^{ij} is used to calculate r^{ij} using (5.10). If there are k feature points in the cluster, then

$$\sigma_r = \sqrt{\frac{\sum_{i=1}^{k}\sum_{j=1}^{k}(r^{ij} - \mu_r)^2}{k}}$$
$$with, \mu_r = \frac{\sum_{i=1}^{k}\sum_{j=1}^{k} r^{ij}}{k} \tag{5.11}$$

denotes the standard deviation of the ratio r^{ij}. A plot of σ_r against n_{in} is shown in Fig. 5.11. The standard deviation of the ratios $r^{ij}, \forall i, j$ decreases as n_{in} is increased, i.e. the ratios become uniform. Thus the distribution of the feature points in lower dimensional feature space can be made a more accurate representation of the distribution in the higher dimensional feature space by using multiple RPMs and then averaging.

The effect of using multiple instances of RPMs and subsequent averaging on the δ values is also investigated. A set of 10 test feature points is randomly distributed in the $\mathbb{R}^{18}$ feature space mentioned above such that δ is zero for all of them. Thus ideally $\delta = 0$ should be preserved for each of these test points in the lower dimensional projected space $\mathbb{R}^2$ as well. If a test feature point lies in the periphery of the cluster formed by the training feature points in $\mathbb{R}^{18}$, there is a chance that in $\mathbb{R}^2$, $\delta \neq 0$ for some of these points because of noise introduced by the RPM. The effect of averaging the values of δ on noise reduction can be explained with the help of observations made in Table 5.2. All the feature points(both from the training cluster as well as the test feature points) in $\mathbb{R}^{18}$ are projected onto $\mathbb{R}^2$ using n_{in} different random projection matrices. Let the δ value obtained for the $p^{th}, p = 1, 2, ...10$ test point in the $q^{th}, q = 1, 2, ...n_{in}$

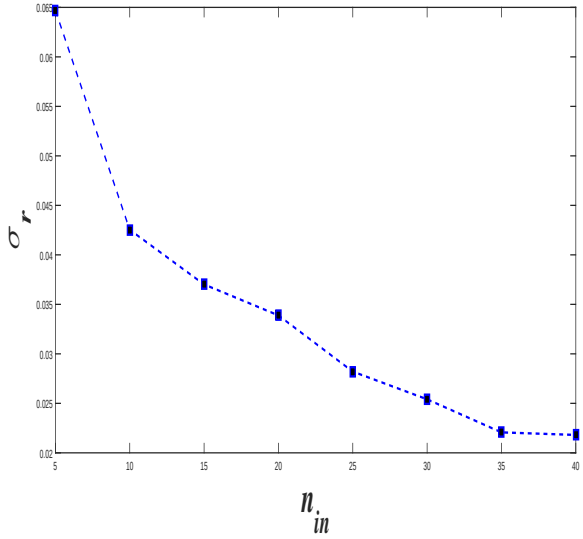


Figure 5.11: Plot of σ_r against n_{in}.

Table 5.2: δ_p values as a function of averaging over different number of instances(n_{in}).

	p = 1	**p = 2**	**p = 3**	**p = 4**	**p = 5**	**p = 6**	**p = 7**	**p = 8**	**p = 9**	**p = 10**
$\boldsymbol{n_{in} = 5}$	0	0	0	0	0	0	0	0	0	0.0511
$\boldsymbol{n_{in} = 10}$	0	0.0504	0	0	0	0	0	0	0.0154	0
$\boldsymbol{n_{in} = 15}$	0	0	0	0	0	0	0	0.0070	0	0
$\boldsymbol{n_{in} = 20}$	0	0	0.0052	0	0	0	0	0	0	0

instance be denoted by δ_{pq}. The mean value of δ for the p^{th} test point over n_{in} instances is given by

$$\delta_p = \frac{\sum_{r=1}^{n_{in}} \delta_{pq}}{n_{in}} \tag{5.12}$$

Table 5.2 tabulates the values of δ_p for different values of n_{in}. It can be observed that for a given value of n_{in} most of the test feature points have $\delta_p = 0$ even in $\mathbb{R}^2$, except for a very few ones. Within these very few points for which $\delta_p \neq 0$, the deviation of δ_p from zero decreases with increasing n_{in}, and this decrease in value saturates beyond $n_{in} = 20$. Thus, for a given set of data points lying in a high dimensional feature space, if we project it into a lower dimensional feature space using multiple number of random projection matrices and average the results, the noise introduced due to this projection gets reduced. The reduction in noise increases with increase in the number of n_{in}.

In order to reduce the distortion introduced by these matrices, the metric $r_{interior}$ is evaluated for a number of different instances of RPMs and the results are averaged. Thus, if $\tilde{r}_{interior}$ denotes the mean of $r_{interior}$ over different instances of RPMs, we define another metric $\tilde{\Delta}$ which gives a measure of how deep within the cluster of training nodes the test node lies.

$$\tilde{\Delta} = \Delta \times \tilde{r}_{interior} \tag{5.13}$$

In a binary classification problem, if the test node is an inlier with respect to both the classes, then the metric $\tilde{\Delta}$ is evaluated for both the classes and compared; in all other cases the metric Δ is compared to obtain the class label CL. The test feature point is assigned the class label with lower value of $\tilde{\Delta}/\Delta$. In the context of the above discussion, a test point is treated as an inlier if it satisfies two conditions:

$$\begin{aligned} \delta &= 0 \\ d_{T,min} &< \max_i d_i^{NN} \end{aligned} \tag{5.14}$$

where $d_{T,min}$ is as defined in (5.8) and d_i^{NN} is the nearest neighbour distance of the i^{th} feature point in the training cluster. For the multi-class problem the extension is as discussed next. Let n_{C_i} denote the number of clusters obtained for class C_i using the methodology described in Section 5.2.1. During the testing phase, let $\Delta_{C_i}^j$ denote the metric obtained using(5.8) for the j^{th} cluster of class C_i. Then the metric

$$\Delta_{C_i} = \min_j \Delta_{C_i}^j, \tag{5.15}$$

$$j = 1, 2, ...n_{C_i} \tag{5.16}$$

$$i = 1, 2, ...M$$

corresponds to the proximity metric of the test feature point to the closest cluster in class C_i. For each of the M classes, the test feature point is examined to determine whether it is an inlier(in the context of satisfying the conditions in (5.14)) or not with respect to its closest

cluster in a particular class. If the test feature point is an inlier with respect to one class only or it does not turn out to be an inlier with respect to any of the classes, then the test feature point is assigned a class label CL according to

$$CL = \underset{i}{\operatorname{argmin}} \Delta_{C_i}, \quad i = 1, 2, ...M \tag{5.17}$$

On the other hand, if the test feature point turns out to be an inlier with respect to two or more classes, then $\tilde{\Delta}_{C_i}$ is evaluated for each of these classes and the class label CL is assigned according to

$$CL = \underset{i}{\operatorname{argmin}} \tilde{\Delta}_{C_i}, \quad i = 1, 2, ...M \tag{5.18}$$

5.4 Modified Feature Sets

In addition to the colour and texture attributes described in Chapter 3, some new attributes have been developed to make the feature sets more robust. To avoid interference from the background region containing part of the pig's face and the surroundings, the muzzle region is segmented out from the background using the method described in Chapter 2. The modified sets of colour and texture features are discussed in the following subsections.

5.4.1 Colour features

Colour feature extraction starts by scanning the pixels which satisfy the following criteria:

- The pixels should lie within the segmented muzzle region.
- The pixels lying within the muzzle region but not carrying any details of the muzzle surface, which can be identified from the Gradient Significance Map(GSM) described in [8].

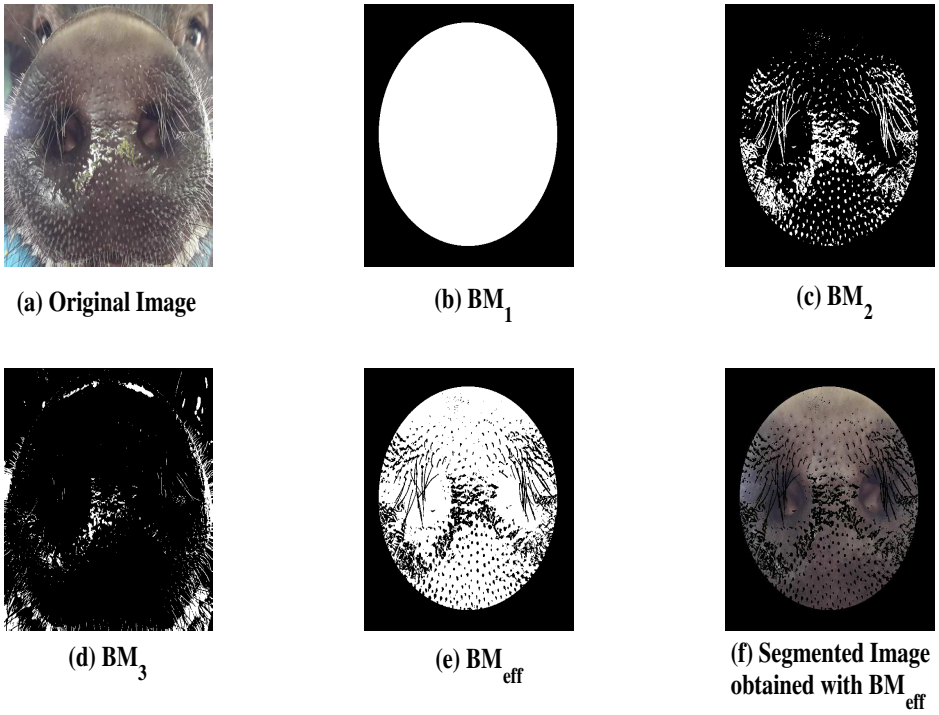


Figure 5.12: The different binary maps generated and the corresponding effective mask for selecting the pixels to be used for colour feature extraction.

- Pixels with intensity component above a threshold(set to 240 in our case for intensities taking values in range 0 to 255) are not considered as these pixels are affected by specular reflection.

An effective binary mask BM_{eff} as shown in Fig. 5.12 is generated fulfilling the above three criteria to select the pixels from the muzzle surface from which the colour information is to be extracted. For a muzzle image, if BM_1 denotes the binary segmentation map obtained using the procedure mentioned in [53], BM_2 denotes the GSM and BM_3 denotes the specularity map, then BM_{eff} is obtained as

$$BM_{eff} = BM_1 \times (F - BM_2) \times (F - BM_3) \tag{5.19}$$

where F is a binary 2D map of the same size as BM_1, BM_2 and BM_3, but with all the elements equal to 1. The multiplication is done on a pixel by pixel basis.

Two sets of colour features have been used which are

- The first set of colour feature f_{C1} consists of μ_{C_b} and μ_{C_r} as the attribute. These attributes have been discussed and derived in Chapter 3 and represent the mean of the C_b and C_r components of all the pixels on the segmented muzzle image. The use of this colour feature set is motivated by the fact that μ_{C_b} and μ_{C_r} should be distinct for each of the four breeds. The values of μ_{C_b} and μ_{C_r} should be more prominently distinct for Hampshires and Yorkshires because Yorkshire muzzle images are completely pink while Hampshires are partially pink in nature. Coming to Duroc and Ghungroo, the muzzle images of most of Duroc are powdery black while some are dual colored like Hampshires. Ghungroo muzzle images are mostly greyish black in nature. Thus

$$f_{C1} = [\mu_{C_b}\ \mu_{C_r}] \tag{5.20}$$

- The second set of colour feature f_{C2} consists of a single attribute only and it is the ratio of pink area to the entire area within the segmented muzzle region. For Yorkshire muzzle images, ideally this ratio should be equal to 1. For Hampshires, this fraction should be between 0 and 1, (ranges mostly from 0.2 to 0.75); for Ghungroo muzzle images this ratio should ideally tend to zero, because Ghungroo muzzle images are completely gray in colour. In the case of Duroc, most of the pigs have a powdery black muzzle with a very small pink region(less than 10% of the segmented muzzle region), while a few of them have a pronounced pink region like Hampshire.

The pixels which are identified by the effective binary mask, need to be labelled as either grey or pink. For pixels belonging to the pink region, the R component in the $[R\,G\,B]$colour vector is greater than the other two. This property is used for a rough identification of the pink region. For the i^{th} pixel if the colour vector is $[R_i\,G_i\,B_i]$, a metric is proposed for this pixel as

$$M_i = R_i - \frac{G_i + B_i}{2} \tag{5.21}$$

If $M_i > TH$, where TH is a pre-set threshold; then the i^{th} pixel is labelled as pink(P), else grey(G). The $[R\,G\,B]$ vector for these labelled pixels is transformed into the YCbCr space. This

transformation is done in order to decouple the luminance and the chrominance information of each pixel. In the YCbCr space, the Cb and Cr components i.e. (C_b, C_r) completely represent the colour information. Thus, the vector $[R_i\, G_i\, B_i]$ for the i^{th} pixel in the RGB space is transformed into $\kappa_i = [Y_i\, C_{b_i}\, C_{r_i}]$ in the YCbCr space. The truncated vector $C_i = [C_{b_i}\, C_{r_i}]$ suffices for our purpose since we do not need the luminance information. Each such colour vector C_i has the associated label $L_i \in \{P, G\}$.Such labelled colour vectors are collected from a set of training muzzle images across all the four breeds. Two separate bivariate Gaussian distributions $G_{\mu_G, \sum_G}$ and $G_{\mu_P, \sum_P}$ are learnt from the data corresponding to the two classes labelled $\{P, G\}$. Here μ and σ are the mean and covariance matrix of the bivariate Gaussian distribution. When a query muzzle image comes, each of the pixels encompassed by the effective mask are tested using a maximum likelihood ratio. Thus if C_j represents a colour vector from a query muzzle image, then the ratio

$$\rho_j = \frac{G_{\mu_P, \sum_P}(C_j)}{G_{\mu_G, \sum_G}(C_j)} \tag{5.22}$$

decides as to whether L_j is to be assigned P or G.

$$L_j = \begin{cases} P, if\, \rho_j > 1 \\ G, if\, \rho_j \leq 1 \end{cases} \tag{5.23}$$

Let n_{pink} denote the number of pixels with $L_j = P$ and n_{grey} denote the number of pixels with $L_j = G$. Then the fraction of pink area within the segmented muzzle η_{muzzle}, which forms our colour attribute for f_{C2} is defined as

$$f_{C2} = \eta_{muzzle} = \frac{n_{pink}}{n_{pink} + n_{grey}} \tag{5.24}$$

The effective colour feature is formed as

$$f_C = [f_{C1}\, f_{C2}] \tag{5.25}$$

5.4.2 Texture features

The texture feature vector defined in (3.21) and denoted as f_{TEX} which is based on the GSM and the Morphological Tophat operator as discussed earlier in Chapter 3 have also been used here. Two more texture feature sets have been proposed in addition to this. The first feature set involves computing the proportion of gray region which is covered by textural details and is denoted as $f_{GRAYDENSITY}$. If n_S represents the number of pixels with structural details in the gray region of the muzzle and n_G represents the number of pixels in the gray region of the muzzle, then

$$f_{GRAYDENSITY} = [\frac{n_S}{n_G}] \tag{5.26}$$

The number of pixels with structural details is obtained from the GSM and the gray region of the muzzle is obtained using the method described in Subsection 5.4.1.

The second feature set involves finding out the amount of non-textural/uniform areas in the muzzle region and is denoted as $f_{UNIFORM}$. The GSM is divided into equal sized non-overlapping patches of size 32×32. The proportion of non-foreground pixels(which represents the uniform areas) within each such patch is computed which are termed as the patch statistic p_s, $0 \leq p_s \leq 1$. With a patch size of 32×32 for a 512×512 image, there are a total of $16 \times 16 = 256$ patches and their corresponding patch statistic. A histogram of 100 bins and uniform bin-width of 0.01 is constructed with these patch statistic values. Let pc_{25}, pc_{50} and pc_{75} denote the $25^{th}, 50^{th}$ and 75^{th} percentile of the patch statistics data as obtained from the histogram. Then $f_{UNIFORM}$ is defined as

$$f_{UNIFORM} = [\mu_{25}\, \mu_{50}\, \mu_{75}\, \mu_{100}] \tag{5.27}$$

where, μ_{25} is the mean of all the patch statistic values between 0 and pc_{25}, μ_{50} is the mean of all the patch statistic values between pc_{25} and pc_{50}, μ_{75} is the mean of all the patch statistic values between pc_{50} and pc_{75} and μ_{100} is the mean of all the patch statistic values between pc_{75} and 1. The effective texture feature vector is then formed as,

$$f_T = [f_{TEX}\ f_{GRAYDENSITY}\ f_{UNIFORM}] \tag{5.28}$$

The colour and texture features derived in this section are fed as input to the MaxST based classifier described in Section 5.2.

5.5 Experiments and Results

The performance of the multi-class classification scheme developed in Section 5.3 is evaluated on the pig breed classification problem using muzzle images and the results are compared with some state-of-the-art multi-class classifiers. The dataset used and the scheme used for splitting the dataset into training and testing parts are the same as discussed in Section 4.3.1. The effective colour and texture features derived in Chapter 3 are concatenated to obtain the composite feature point f^{MUZZLE}

$$f^{MUZZLE} = [f_C f_T]; \tag{5.29}$$

The calculation of $\Delta_{C_i}^j$ for $j = 1, 2, ...n_{C_i}, i = 1, 2, 3, 4$ using (5.8)requires the involvement of two matrices A_S and $B_{S_{EXT}}$. The computation of the matrix $B_{S_{EXT}}$ requires the test feature point and hence its calculation is done in the testing phase. On the other hand the computation of the matrix A_S requires only the training feature points. Let A_S^{ij} denote the adjacency matrix obtained for the j^{th} cluster of the i^{th} class. So the total number of adjacency matrices to be learnt in the training phase is

$$N = \sum_{i=1}^{M} n_{C_i} \tag{5.30}$$

If a test feature point satisfies the conditions of (5.14) for two or more classes, then $\tilde{\Delta}$ needs to be evaluated for these classes and the class label CL is assigned according to (5.18). To evaluate $\tilde{\Delta}$ according to (5.13), first $\tilde{r}_{interior}$ needs to be computed. The computation of $r_{interior}$ using (5.9) involves projecting the training feature point as well as the test feature point to the $\mathbb{R}^2$ space using random projection matrices. The number of instances n_{in} over which $r_{interior}$ is to be computed to obtain $\tilde{r}_{interior}$ is very crucial from the point of view of

Table 5.3: Mean classification accuracy for 100 iterations and their standard deviations for different values of n_{in}.

	$n_{in} = 5$	$n_{in} = 10$	$n_{in} = 15$	$n_{in} = 20$
Duroc	90.61(7.28)	91.32(7.31)	92.44(7.37)	93.26(7.22)
Ghungroo	96.42(5.37)	96.98(5.09)	97.16(4.88)	97.19(4.35)
Hampshire	90.87(11.07)	91.73(10.46)	92.78(10.58)	93.61(10.21)
Yorkshire	100(0)	100(0)	100(0)	100(0)

Table 5.4: Mean classification accuracy for 100 iterations and their standard deviations.

	Proposed method	**Our earlier approach in [5]**	**SVM (Gaussian kernel)**	**k-NN (k = 10)**	**Decision Trees**
Duroc	93.26(7.22)	86.45(12.05)	86.96(9.92)	83.66(12.49)	77.46(16.61)
Ghungroo	97.19(4.35)	93.02(7.46)	93.95(8.05)	95.2(6.92)	89.93(11.75)
Hampshire	93.61(10.21)	86.91(12.51)	89.54(12.74)	85.9(13.79)	76.81(15.28)
Yorkshire	100(0)	98.54(4.84)	99.88(0.91)	100(0)	95.68(5.43)

Table 5.5: Confusion matrix for proposed algorithm.

Actual \ **Predicted**	**Duroc**	**Ghungroo**	**Hampshire**	**Yorkshire**
Duroc	66	3	2	0
Ghungroo	3	83	0	0
Hampshire	3	1	67	1
Yorkshire	0	0	0	119

minimizing the noise induced by the random projection matrix to the data distribution after projection in $\mathbb{R}^2$ space. The impact of the choice of n_{in} on the classification accuracy are shown in Table 5.3. The training and testing were carried out for 100 iterations with different random combinations of training and testing muzzle images. Completely separate sets of pigs were used in training and testing. The mean classification accuracies along with their standard deviations(in parenthesis) over 100 iterations, are listed in Table 5.3. It can be observed from the Table that the classification accuracies tend to improve with increasing n_{in} which corroborate with the observations made in Table 5.2. These accuracies saturate beyond $n_{in} = 20$. The improvement in the classification accuracies with increasing n_{in} clearly shows the significance of using multiple RPMs and averaging the results in the lower dimensional space as discussed in Section 5.3.

Table 5.4 compares the classification accuracies obtained using our method with some state-of-the-art classifiers in literature. The superiority of the proposed classifier can be observed from the results in Table 5.4. The proposed classifier combined with the refined set of features gives better classification results as compared to what was obtained using the hierarchical classification scheme listed in Table 4.9. The confusion matrix obtained using the proposed method is also shown in Table 5.5. From the table it can be observed that the maximum confusion arise between two pair of breeds which are: Duroc/Ghungroo and Duroc/Hampshire. Duroc/Ghungroo confusion comes from texture(a very few Duroc pigs have texture density similar to that of Ghungroo) as well as colour front(muzzle images of both breeds are mostly shades of grey). The other confusion between Duroc/Hampshire comes from the fact that both the breeds contain some pigs which have identical fraction of pink region in their muzzle images.

6

Classification Scheme based on Outlier Detection Framework Using Maximum and Minimum Spanning Trees

Contents

As discussed in chapter 5, based on the MaxST representation learnt for a particular cluster of a given target class, a test feature point is assigned an outlier score during the testing phase. It was observed that when the test feature point T lies inside the cluster, the metric $\delta(T)$ in (5.2) evaluates to zero, irrespective of the location of the test point inside the cluster. Thus, when this outlier detection framework is applied to make inference about the class label of a test feature point T which lies in the overlapping region of two or more classes in a multi-class classification problem as discussed in Section 5.3, a metric $r_{interior}$ was proposed which provides an estimate of how deep within the cluster, a test feature point lies. There are two issues in the computation of the metric $r_{interior}$ using (5.9):

- The computation of the metric $r_{interior}$ using (5.9) involves checking the number of instances over $\binom{k}{3}$ triangles for which the test feature point lies inside, where k is the number of training feature points within the cluster. For even small to moderate values of k, the computational complexity becomes very high.
- The projected data in $\mathbb{R}^2$ is affected by noise due to the use of Random Projection Matrices as discussed in Section 5.3.

In order to avoid these two issues, a new formulation is proposed based on both MaxST and Minimum Spanning Trees(MinST). The MinST is used to assign outlier scores to the test feature point based on its positional arrangement with respect to the feature points of the training cluster, when the test feature point lies within the cluster.

6.1 Outlier Detection Methods in literature

Given a cluster of data-points, it is important to first build a stable, robust and compact model for cluster. When constructing the natural inlier boundary surface covering most of the training points, the construction procedure must draw a line between extreme data-sensitivity and over-generalization. If the boundary-surface is data-sensitive, then the model will not be stable, since not all training points can be considered typical model-representatives (some of them can be noisy inlier points atypical of the model being represented).

6. Classification Scheme based on Outlier Detection Framework Using Maximum and Minimum Spanning Trees

Extreme data-sensitivity, is typically the problem with localized outlier detection frames [80–82], where the test data-point is placed along with the training set/cluster and a conformity check is done about the test-point location. This local conformity check of the test data-point can be performed in terms of parameters such as number of nearest neighbours [83], point density profile [80], degree of connectivity [81] etc.

Minimum spanning trees (MinSTs) are useful in binding data-points which have a high correlation or statistical similarity. For datasets wherein the data is either limited and/or the clusters have an arbitrary shape, MinSTs can be used to generate the back-bone or skeleton about which the shape of the cluster coupled with the tolerance bands can be produced [64]. In Juszczak et al. [64] the authors proposed a method named MST-CD in which the MinST was used as a cluster-descriptor for the target class. Once the cluster-model was generated, the distance of the test data-point from the closest edge in the MinST was chosen as the outlier score.

Localized pattern checking works best when the training data exhibits some form of a regularized pattern over space, not necessarily completely homogeneous, but with some variability which can be tracked and profiled. This clearly does not mean that if one positions oneself at a particular data-point in space, the statistical view will be the same and independent of the data-point picked. This means that when positioned at a particular point, its statistical view is likely to be similar to its k-nearest neighbours. Beyond this there will be a change in the statistical profile, whether this happens to be in the form of a density map or some other form of crude statistic based on the k-nearest neighbourhood distance profile.

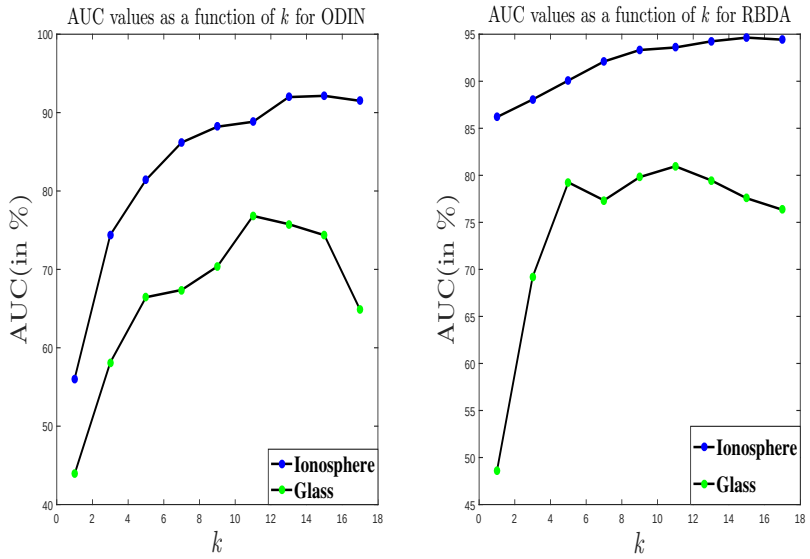
Among the unsupervised methods, the most classical one is the $k - NN$ which uses the distance of the k^{th} nearest neighbour as the outlier score. Such a method also purely relies only on distance of the nearest neighbours. An example from literature is the Local Outlier Factor (LOF) which considers the ratio of the density of a test point to the density of its k neighbouring points and assigns an outlier score based on this ratio. In [84] the authors demonstrate a case in which the efficacy of the LOF will reduce when the density of the neighbouring points is close to the density of an outlier test point. Among other popular methods based on nearest

neighbours are the ODIN [81] and RBDA [82].

In ODIN [81], the scores are assigned based on the number of instances for which a test point lies within the k-nearest neighbours of the remaining data. In other words the degree of connectivity of the test point with its neighbours, is established based on similarity in certain local statistical parameters such as density etc. In RBDA [82], the average proximity rank of the test-point is produced with respect to its k nearest neighbours, considering the distance profiles of neighbours of these immediate neighbours. This is an indirect spatial continuity check and in subtle way indicates whether the inclusion of the test point has changed anything at all with respect to the local distance profile pattern (about the test point).

Converging at the best possible statistic to measure statistical conformity of a test point about a local neighborhood, has been the motto of most of these local outlier detection frames. Arriving at first set of k nearest neighbours based on a minimum distance threshold is essentially a catch-22 problem [81]. These nearest neighbour methods do not consider the distribution of the entire data from target class (i.e. uni-modal or multi-modal, the bigger picture is usually lost). These methods are sensitive to the selection of the parameter k. Fig. 6.1 shows the variation of the performance metric AUC-ROC of the two nearest neighbour algorithms ODIN and RBDA. It can be observed that for the "Ionosphere" dataset the AUC-ROC value improves with increasing k for both the methods and finally saturates to a constant value. However, for the "Glass" dataset the trend is completely different. The performance improves with increasing k till it reaches an optimum value, beyond which the performance again falls. Thus the value of k which is the most crucial parameter of nearest neighbour algorithms has to be adjusted depending upon the nature of the dataset.

There always remains an uncertainty whether the right and pertinent neighbours have been chosen in this primary set for further analysis. Thus, a pre-screening strategy was introduced by Tang et al. [80], wherein, data points which belonged to either the k nearest neighbours or k reverse nearest neighbours or k shared nearest neighbours were used to form the primary local-set over which the conformity check was to be performed. A reliability check was done to find out whether the subsequent outlier detection process resulted in a detection of a test



mework

Figure 6.1: Variation of AUC-ROC values as a function of the number of nearest neighbours k for two different nearest neighbour methods ODIN and RBDA.

point within a compact spherical region encompassing the three different types of neighbours used for this analysis. While conformity checks are useful, they assume the datasets to be largely homogeneous in space with a slight variability. For datasets (e.g. "Glass" and "Musk" as discussed in Section 6.3), which contain several micro-clusters and are of a sparse and heterogeneous type, these methods may not be that effective.

In another variant, instead of using a single value of k (k, now a locally variable parameter), the authors in [85] propose a method called TOD in which the information regarding the nearest neighbour distances obtained for multiple values of k are combined. For each instance of k, the product of k^{th} nearest neighbour distances of two data points is computed for all possible pairs and stored in a matrix. The information regarding the deviation in this product across the range of k is combined for all the pairs of data points (i, j). Finally, by choosing triplets of data points and based on a triangle inequality theorem, the points with the highest nearest neighbour distances across a range of values of k is found out and assigned as outliers. However, the lack of a systematic procedure for the proper choice of the range of k values along with the complexity of the method results in a poor performance on many standard datasets.

In Danesh et al. [86], the primary data was assumed to be a composition of multiple smaller

clusters and a combination of both labelled and unlabelled samples. To detect the most negative samples which are unlikely to be a part of the main model, an outlier detection scheme named EODSP was presented. For all the unlabelled samples, the posteriori probability that the sample belongs to one of the classes, say C_i, was computed based on a normalized relative distance from that class centroid from C_i. For each unlabelled sample an entropy measure was computed over all the classes with the class-posteriors as a base. Any skew in the conditionals would be indicative of a cluster proximity and hence, the test-sample is likely to be an inlier. Thus, higher entropy scores are likely to indicate that the test-point is most likely an outlier. The problem in using this information for outlier detection is that micro-details are lost as the computations are global. This approach will also fail for cases where the data can be properly split into several compact clusters (i.e. for datasets such as "Musk").

Data pre-processing and massaging helps in streamlining the training data so that irrelevant details are ironed out and the subsequent model remains robust to intra-class variations. Furthermore, this channelization helps in detecting outliers more reliably. There are some datasets such as "Breast(diagnostic)", wherein the number of attributes or the dimensionality of the data $d = O(n)$, where, n is the number of training samples. Under these circumstances, the problem demands a compaction of the feature vector by removing attributes which are redundant and where either the variability [87] or dependence [88] is minimal. This can be done via a variability check using an eigen-space analysis [87].

In Riahi et al. [89], a method named SODEP was proposed, which used Principal Component Analysis to divide the feature space into three different sub-spaces, on the basis of the variance associated with each principal component, outlier scores are evaluated in each of these sub-spaces and finally they are combined to generate an overall outlier score. In [87], another method named mRMRD is proposed in which the subspace is formed from the dimensions along which the density of the data varies the most. The criterion used to select a particular dimension is based on the mutual information between the distribution of data density in feature space and along that particular dimension. A particular dimension is retained in the subspace if it has a sufficiently high mutual information. Owing to the truncation of some of the

attributes, coarse structural information is lost and original shape of the cluster (if monolithic) is compromised to some degree. This affects the performance of sub-space based methods particularly for datasets where the data is heterogeneous and sparse (i.e. where the clusters are not well defined), e.g. "Glass" and "Musk".

6.2 Proposed Framework for Outlier Detection

The proposed outlier detection framework is an extension of the method described in Chapter 5 which uses a MaxST formulation. In the current framework, we use a MinST in conjunction with a MaxST. While the MaxST is more sensitive in terms of the structural change induced by the test feature point T when T lies outside the cluster, the functionality of the MinST is just the opposite i.e. the structural change in the MinST depends on the location of T within the cluster. Thus the functionality of the MaxST and MinST complement each other in assigning outlier score to the test point T. The training process of the proposed outlier detection framework is similar to the one outlined in Section 5.2, which involves first splitting the training data from the target class into multiple compact clusters and then learning MaxST representation for each such cluster. In the current framework, both MaxST and MinST representations are learnt for each of the compact clusters into which the training data has been split into. The process of splitting the training data into multiple compact clusters, the learning of MaxST representation and the associated testing scheme has been described in detail in Section 5.2.2 respectively. The learning of MinST representation along the associated testing scheme and its efficacy are described next.

6.2.1 Score generation for test-points in the cluster interior based on MinST analysis

When the test-point falls inside the cluster (or within the convex hull), the MaxST δ score ends up being zero. There is therefore a need to modify the scoring to assess how deeply the test-point is embedded inside the cluster. This assessment is done via a MinST analysis. Hence, what cannot be derived from the MaxST analysis is provided by the MinST analysis. Thus two

graphical models, one based on MaxST and the other based on MinST are required to cover both scoring-angles related to the exterior of the convex hull and the interior respectively.

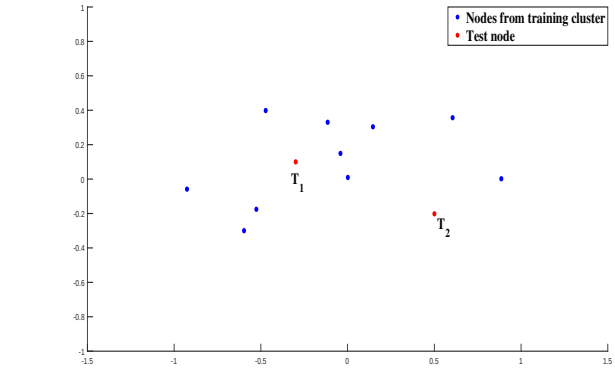


Figure 6.2: Two test nodes with different positional arrangements with respect to the nodes in G.

As observed from the discussions in Section 5.2.2 and Fig. 5.4, for a test node T, which lies either inside the convex hull of the cluster or in the vicinity of the cluster, the metric $\delta_{MAX}(T)$ evaluates to zero. In such a case depending upon the positional arrangement of the test node with respect to the nodes of the training cluster, it's probability of being an outlier with respect to the nodes of S will change. For instance, consider the two different positional arrangements corresponding to two test nodes T_1 and T_2 with respect to a set of nodes G from a training cluster as shown in Fig. 6.2. Although, both $\delta(T_1)$ and $\delta(T_2)$ evaluate to zero it is quite evident from Fig. 6.2 that test node T_1 belongs more to the cluster G as compared to T_2. That is T_2 should have a higher outlier score as compared to T_1. Thus the metric $\Delta(T)$ can be modified further to include the effect of this positional arrangement of the test node T.

It has been observed that in the formulation of $\delta_{MAX}(T)$ described in Subsection 5.2.2, if the MaxST is replaced by a MinST, then the information regarding the positional arrangement of T can be within the convex hull can be captured. This observation is shown in Fig. 6.3 for the two test nodes T_1 and T_2. Here, the notations $MinST_{ORIG}$ and $MinST_{EXT}$ have the same meanings as their MaxST counterpart in Section 5.2.2. Since, T_1 is surrounded by many neighbours from G, it induces a higher change in the MinST structure. As a result, it registers a higher value of $\delta = 0.6167$ as compared to $\delta = 0$ for T_2. Let $\delta_{MIN}(T)$ denote the counterpart of $\delta_{MAX}(T)$

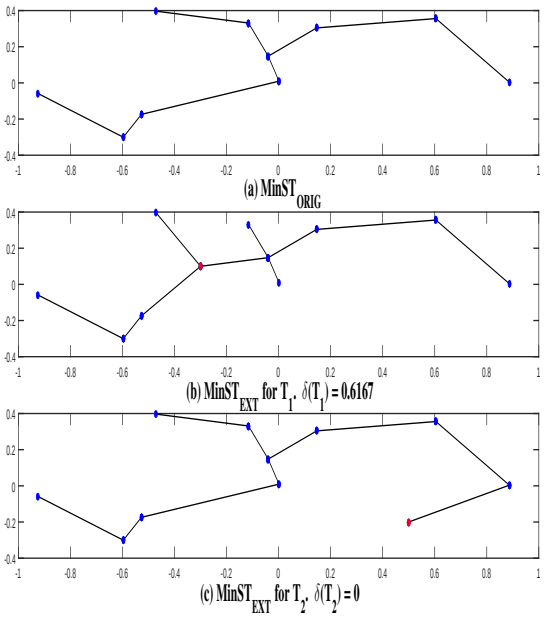


Figure 6.3: Two test nodes with different positional arrangements with respect to the nodes in G.

in (Eqn. 5.2) using the MinST construct. Then, the score positioned as an inlier-outlier with respect to the k^{th} cluster is given by $S_{MIN_{T(k)}}(T)$ and is defined as,

$$S_{MIN_{T(k)}}(T) = 1 - \delta_{MIN}(T)$$
$$where, \delta_{MIN}(T) = \frac{||A_S^{MIN} - B_{S_{EXT}}^{MIN}||_F}{||A_S^{MIN}||_F} \tag{6.1}$$

where the matrices A_S^{MIN} and $B_{S_{EXT}}^{MIN}$ also have the same meaning as their MaxST counterpart. Similar to the metric $\delta_{MAX}(T)$ derived in (Eqn. 5.2), $\delta_{MIN}(T)$ also satisfies the property $0 \leq \delta_{MIN}(T) \leq 1$. With this formulation, greater the change induced by the induction of the test point in the MinST, larger will be the δ_{MIN} score. Hence, deeply embedded test-points in denser zones are likely to generate higher values of δ_{MIN}.

6.2.2 Overall Outlier Score from MaxST and MinST analysis

Given the training data from the target class which contains multiple compact secondary clusters as identified by the method described in Section 5.2.1, the outlier scores are gener-

ated for each of these clusters. Let M denote the number of such clusters and $S_{MAX_T(k)}$ and $S_{MIN_T(k)}, k = 1, 2, ..., M$ denote the MaxST and MinST outlier scores respectively generated for each such cluster, with respect to a particular test-point T. Then the outlier score for the target class with respect to either the MaxST or MinST formulation is decided by the outlier score of T corresponding to its closest cluster in the training data. Thus if $SMIN_{MAX_T}(T)$ denotes the outlier score for the target class with respect to the MaxST formulation, then

$$SMIN_{MAX_T}(T) = \min_{k \in 1,2,..M} S_{MAX_T(k)}(T)$$
$$where, S_{MAX_T(k)}(T) = \delta(T) \tag{6.2}$$

$\delta(T)$ being defined according to (5.2). Similarly, if $SMIN_{MIN_T}(T)$ denotes the outlier score for the target class with respect to the MinST formulation, then

$$SMIN_{MIN_T}(T) = \min_{k \in 1,2,..M} S_{MIN_T(k)}(T) \tag{6.3}$$

where $S_{MIN_T(k)}(T)$ is defined according to (6.1). The overall outlier score can be written as,

$$S_{OUTLIER} = (SMIN_{MIN_T}(T) + SMIN_{MAX_T}(T)) \times d_{NN}(p, T) \tag{6.4}$$

where, $d_{NN}(p, T)$ is the mean of the p-nearest neighbour distances to test-point T and $SMIN_{MAX_T}(T)$ and $SMIN_{MIN_T}(T)$ are given by Equations Eqn. 6.2 and Eqn. 6.3 described earlier. Including the nearest neighbour distance in the formulation is extremely important since both the Overall-MaxST and the Overall-MinST normalized scores saturate beyond a particular distance. At that point the nearest neighbour distance becomes necessary, as an absolute spatial reference.

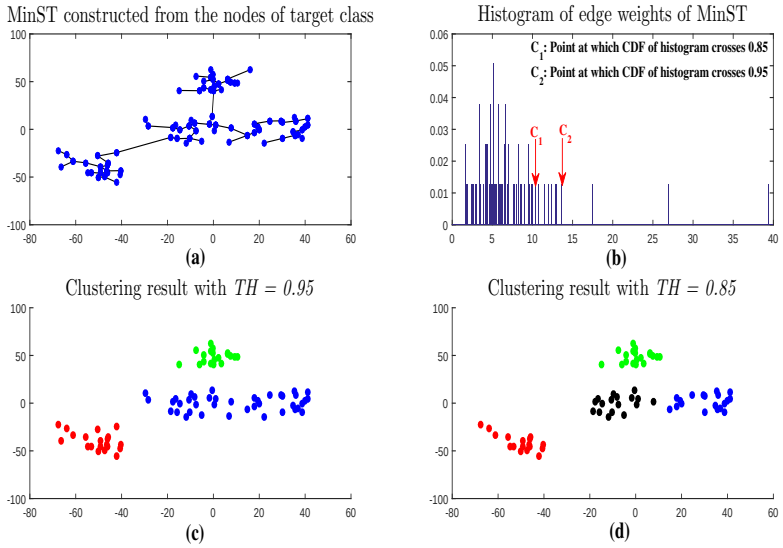


Figure 6.4: Effect of TH on the number of fragments generated.

6.3 Experiments and Results

The performance of the proposed outlier detection framework is mainly dependent on two parameters. The first parameter is the threshold TH described in Section 5.2.1 which controls the number of fragments into which the target training data is split. To illustrate the effect of the parameter TH on the number of fragments generated, consider the synthetic dataset in Fig. 6.4.

Fig. 6.4(a) shows the nodes from a target class distributed in feature space and the MinST constructed out of these nodes. The histogram constructed out of the edge weights in the MinST is shown in Fig. 6.4(b). It can be observed that there are long as well as short string of nulls in this histogram. Two different points in the histogram denoted as C_1 and C_2 corresponding to values of $TH = 0.95$ and $TH = 0.85$ are also shown. The different fragments generated using these two values of TH are shown in Fig. 6.4(c) and (d) respectively. Although the set of nodes shown in blue in Fig. 6.4(c) can be treated as a single cluster, this set of nodes is split up into two different clusters as shown in Fig. 6.4(d) when TH is changed to 0.85. This over fragmentation happens because as we reduce the value of TH many more string of prominent nulls participate in deciding the value of the threshold e_T. The effect of over fragmentation

on some real world datasets will be shown when we consider the performance analysis of our outlier detection framework on these real world datasets.

The second parameter which we denote as r_{min} sets a limit on the minimum size of the cluster(after fragmentation); satisfying which it can be considered for further processing. Here r_{min} denotes the minimum fraction of the total number of data points which is required to be present within a cluster after fragmentation. The parameter r_{min} is important because it tends to reject some of the noisy nodes within the training data and hence improves the efficiency of the MaxST representation. In all our experiments, we set $r_{min} = 5\%$.

6.3.1 Performance evaluation of outlier detection framework

The UCI repository [90] consists of several datasets of different sample sizes and dimensions. The multi-class problems are transformed to one-class by selecting one of the classes as the target class and the rest as outlier. 30% of the data from the target class are randomly selected for training and the remaining 70% data form the target class as well as the entire outlier data was kept for testing. Thus our outlier detection framework uses data from the target class alone for training purpose unlike many other outlier detection framework in literature. The Area Under Curve(AUC) obtained from ROC [91] can be used to assess the performance of one class classifiers and is one of the most popular metric for this purpose. Higher values of AUC indicates better performance of one class classifiers. A value of AUC less than 50% indicates that the performance of the one-class classifier is worse than random guessing. In our experiments, the training and testing cycle was carried out for 20 times with different training and testing data and the mean AUC values over 20 iterations have been recorded. The datasets used from the UCI repository for the performance evaluation of our outlier detection framework along with their details are mentioned in Table 6.1.

As can be observed from Table 6.1, the datasets chosen contain both low dimensional as well as high dimensional data. The high dimensionality of a dataset affects the performance of any machine learning algorithm due to the phenomenon of curse of dimensionality [92].There are datasets with different number of training samples as well. The effect of the parameter

Table 6.1: Datasets used from the UCI repository with their descriptions.

Datasets	# of instances from target class	# of outlier instances	Dimension of data
Ionosphere	225	126	33
Glass	205	9	9
Musk	2965	97	30
MNIST	6903	700	166
Breast (Diagnostic)	357	21	100

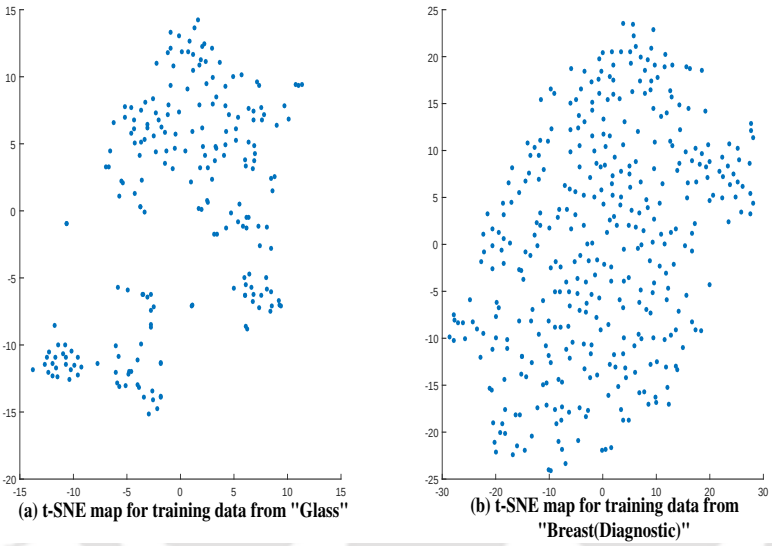


Figure 6.5: Visualisation of the training data from the "Glass" and "Breast(Diagnostic)" datasets in $\mathbb{R}^2$ space.

TH on the AUC values are first observed. Table 6.2 shows the AUC values obtained for the datasets mentioned in Table 6.1 for different values of TH. The AUC values show some change with variation in TH for the datasets "Glass" and "Ionosphere"; whereas for the remaining datasets there is virtually no change in the AUC values with TH. For the datasets "Glass" and "Ionosphere", the AUC values tend to show some improvement with increasing TH. This phenomena of AUC values changing with TH for some of the datasets, while not changing for the others can be attributed to the nature of the dataset. Consider the distribution of training nodes in feature space from two different datasets: "Glass"(for which the AUC values change with TH) and "Breast(Diagnostic)"(for which the AUC values do not change with TH). The

Table 6.2: Variation of AUC values(in %) with TH.

Datasets	**TH = 0.6**	**TH = 0.7**	**TH = 0.8**	**TH = 0.9**	**TH = 0.99**
Ionosphere	94.19	94.75	95.08	96.13	95.89
Glass	85.50	84.43	84.83	86.06	87.24
Musk	99.95	99.95	99.95	99.95	99.95
MNIST	92.75	93.00	92.89	93.02	93.23
Breast (Diagnostic)	94.77	94.49	94.90	95.21	95.03

visualization of these datasets is done in $\mathbb{R}^2$ space by projecting the data in $\mathbb{R}^2$ space using the t-SNE [38] mapping method. These visualizations are shown in Fig. 6.5. The t-SNE maps show that the training nodes from the dataset "Glass" are scattered in multiple clusters in feature space and hence the AUC values are affected by the choice of the parameter TH; whereas for the dataset "Breast(Diagnostic)" the nodes representing the training data tend to be united as a single cluster and thus the fragmentation algorithm described in Section 5.2.1 has negligible effect on the AUC values. For the datasets "Glass" and "Ionosphere" best performance is achieved for $TH \geq 0.9$. One of the most important observations from Table 6.2 is that beyond $TH = 0.9$ there is very little change in the performance of our algorithm across datasets of varying size and dimensions. Thus our algorithm is less sensitive to the parameters TH and r_{min} over a wide range of operation. This is a great advantage of our methodology over other methods which are highly sensitive to the variation in their parameters, especially the class of nearest neighbour algorithms which depend on the number of nearest neighbours chosen as a parameter of the system.

Table 6.3 compares the performance of our outlier detection framework with several other state-of-the art methods from literature in terms of the AUC obtained from the ROC curve. The ROC curves are shown in Fig. 6.6. The performance with the MaxST formulation alone is also shown in this Table. In Table 6.3, method **P1** refers to the AUC values generated from the outlier scores obtained using (6.4); while method **P2** refers to the AUC values generated by excluding the $SMIN_{MIN_T}(\bar{T})$ term from (6.4). Thus, by observing the difference in the AUC values obtained using methods **P1** and **P2**, the importance of considering the

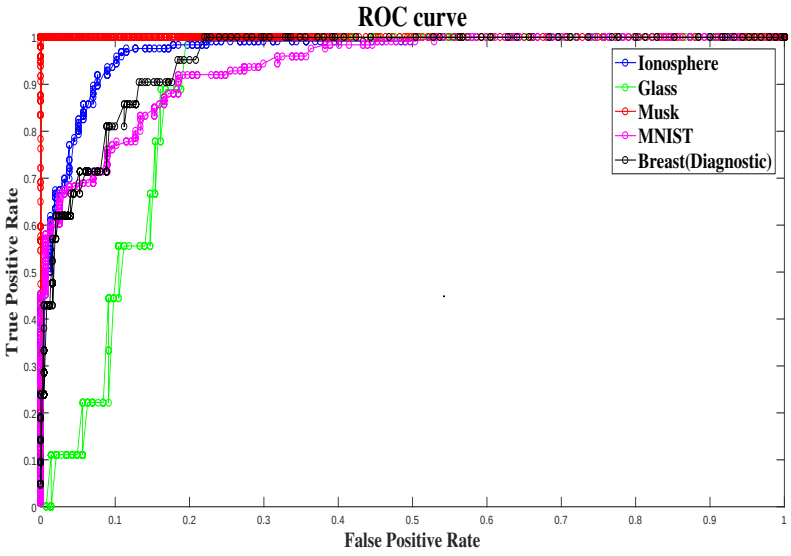


Figure 6.6: ROC curves corresponding to the five datasets mentioned in Table 6.1.

positional arrangement of the test node with respect to the nodes of training cluster can be understood. For all the five datasets, by considering the positional arrangement of the test node with respect to the nodes of the training cluster, an improvement in the performance of our proposed method is observed. This improvement is most significant for the case of "Ionosphere" dataset, where there is a nearly 3% jump in the AUC value. Our proposed outlier detection framework outperforms the others like MST-CD [64], LOF [93], ODIN [81], RDOS [80], TOD [85], EODSP [86], SODEP [89], MRD/mRMRD [87] and HiCS [88] from literature in four out of the five datasets. The performance of our algorithm is better for the high dimensional datasets "MNIST" and "Breast(Diagnostic)" and also for the moderate dimension datasets "Ionosphere" and "Musk", when it is compared to methods like SODEP, MRD/mRMRD and HiCS which tries to select a suitable subspace for high dimensional data and then use existing outlier detection approaches like LOF [93] or k-NN in these selected subspace. The data in the lower dimensional subspace is always an approximation to the high dimensional data irrespective of the algorithm used to select the appropriate subspace for representation. On the contrary, our algorithm does not truncate any attribute of the data. Instead, in order to mitigate the curse of dimensionality, our proposed algorithm tries to split

Table 6.3: Performance comparison of our proposed Outlier Detection framework with other state-of-the-art methods in literature in terms of the AUC values. Here $P1$ refers to the proposed outlier detection framework by using the MinST formulation in conjunction with the MaxST formulation for generating inlier scores whereas $P2$ refers to using only the MaxST formulation for generating the outlier scores.

Datasets	Ionosphere	Glass	Musk	MNIST	Breast (Diagnostic)
MST-CD [64]	91.68	88.89	96.00	86.17	86.53
LOF [93]	90.47	86.82	63.81	80.34	87.41
ODIN [81]	92.14	76.82	98.83	91.21	93.18
RDOS [80]	76.12	65.00	90.27	75.43	88.45
TOD [85]	62.37	57.5	72.35	65.82	85.75
EODSP [86]	88.15	81.93	72.41	86.32	90.49
SODEP [89]	91.41	80.97	92.86	87.77	88.39
MRD/mRMRD [87]	95.18	**92.62**	91.84	88.06	91.51
HiCS [88]	82.34	80.05	89.21	51.74	94.23
P1	**96.13**	87.24	**99.95**	**93.23**	**94.81**
P2	93.04	86.24	**99.95**	92.56	94.55

the data into multiple clusters and then learn appropriate representation for each of them which is a big advantage of our proposed method.

In order to compare the performance of our proposed model with the SVDD [2], we cannot use the AUC obtained from the ROC curve as a performance metric. This is because the SVDD produces crisp outputs in the form of labels, which classify a test point as either inlier or outlier. Hence, for performance comparison with SVDD, we directly compute the classification accuracies; i.e. fraction of the test points for which the computed labels match the ground truth. In order to generate the classification accuracies with our proposed method, we compare the outlier score obtained using (6.4) with a threshold. This threshold value is chosen such that the classification accuracy is maximized. Table 6.4 shows the results. Clearly our method

Table 6.4: Performance comparison of our proposed Outlier Detection framework with Support Vector Data Description(SVDD) [2] in terms of classification accuracy.

Datasets	**Classification accuracy obtained using SVDD (in %)**	**Classification accuracy obtained using proposed framework(in %)**
Ionosphere	83.04	84.12
Glass	81.81	86.48
Musk	64.47	97.64
MNIST	72.65	90.38
Breast (Diagnostic)	94.21	94.83

Table 6.5: Classification accuracies for the four breeds obtained using (6.5). The mean accuracy across 100 iterations with different random selection of training and test data. The values in parenthesis indicate the standard deviation.

Breeds	**Mean Accuracy(in %)**	**St.Dev. of Accuracy(in %)**
Duroc	93.85	7.37
Ghungroo	97.48	4.16
Hampshire	94.27	9.86
Yorkshire	100	0

outperforms the SVDD on all the five datasets.

6.3.2 Performance evaluation for Multi-class Classification

For a multi-class classification problem with N_T number of classes($N_T = 4$ for our breed classification problem), let $S^i_{OUTLIER}, i = 1, 2, ..N_T$ denote the outlier score generated for class i according to (6.4). Since for a test feature point T, a lower value of $S^i_{OUTLIER}$ indicates lower probability of T being an outlier to class i, thus the inference about the class label CL of T is made according to

$$CL = arg\min_i S^i_{OUTLIER} \tag{6.5}$$

The classification accuracies for the four breeds obtained using (6.5) are listed in Table 6.5. The training and testing procedure is the same as discussed in Section 4.3.1, i.e. the set of pigs used for training is completely different from the set used for testing purpose and a

$50\% - 50\%$ split was done for training and testing purpose. Although the mean accuracies for the four breeds across 100 iterations with different training and testing data remain almost the same as those listed in Table 5.4, however the computational complexity of classification is reduced drastically using the current framework. This result also shows the ability of the MinST framework discussed in Section 3.28 in resolving the problem of overlapping classes, i.e. when the test feature point falls in the overlapping region of two or more clusters from different classes.

7

Summary and Future Work

Contents

Visual animal biometrics has gained extensive popularity because of the process being non-intrusive in nature and robust to theft or other fraudulent activities. As compared to individual identification, breed identification has more commercial significance for domestic animals like cows, goats, pigs etc. In this thesis an attempt has been made to classify four breeds of pigs viz. Duroc, Ghungroo, Hampshire and Yorkshire which are of commercial significance, based on their muzzle images. The muzzle region forms both a tactile as well as nasal interface with the surrounding and there are many sensory features present over this region. Since biometric traits generally tend to be concentrated at the sensory interfaces, hence it was concluded that the muzzle image could be as a breed identifier. It was observed that the colour and texture profile of the muzzle image could be used as potential breed identifiers. Robust texture and colour statistics were extracted based on the discriminatory features between the breeds and then customized classification techniques were developed. This chapter concludes the thesis with a summary of the work done and provides future research directions which can further improve the robustness of the breed identification system for pigs. The summary of the entire breed identification process involving muzzle segmentation, feature extraction and the design of classification algorithms are enumerated next.

7.1 Summary

- **Muzzle Segmentation:** The muzzle images to be used for feature extraction contain significant background information, which act as a noise, thus corrupting the feature vectors. Hence it becomes essential to isolate the muzzle region from the background prior to feature extraction. Because of colour and texture profiling issues, the muzzle contour could not be estimated accurately even with state-of-the-art active contour segmentation techniques in literature. Since, it was observed that the texture statistics were highly sensitive to background interference, hence instead of going for an over-estimate of the muzzle contour, an attempt was made to fit a circular ball shaped segmentation mask with the largest possible radius inside the muzzle contour. The segmentation process starts with the application of a 2-dimensional pre-filtering operator on the muzzle image which

uses a Derivative of a Gaussian operator with a smoothing wing along the orthogonal direction. The parameters of this filter are carefully selected in such a manner so as to highlight the muzzle contour while suppressing the textural details in the interior of the muzzle. The response of the filter is then thresholded to obtain a binary map. An algorithm has been developed which selects the foreground pixels in the binary map lying within an annular region and computes the radius of the circular mask based on the distance of all the foreground pixels in this annular region from the centroid of the image. The radius of the circular segmentation mask thus depends heavily on the nature of the muzzle contour. It was observed that there was a significant improvement in the classification accuracies for all the four breeds with this adaptive circular segmentation mask applied, as compared to the case when no segmentation was done.

- **Feature Extraction:** Once the muzzle region has been segmented out using the circular segmentation mask discussed above, the next step is feature extraction. While computing the statistics from the texture and colour descriptors, only pixels falling under the segmentation mask are considered. On the texture front, two descriptors have been put forward: the Gradient Significance Map and the response of the muzzle image to the Morphological Tophat operator. While the Gradient Significance Map highlights the cilia, pores, muzzle contour, portions of the nostril boundary and also the boundary of the pink region within the muzzle in the case of dual-coloured muzzle; the response of the Morphological Tophat operator only highlights the pores and cilia on the muzzle surface. This is useful in discriminating certain breeds such as Ghungroo and Hampshire. The texture feature vectors are extracted by dividing the Gradient Significance Map and the response of the muzzle image to the Morphological Tophat operator into four quadrants and extracting suitable statistic from each such quadrant. On the colour front, the image was analysed in the YCbCr domain to extract the colour feature vectors. This was done in order to decouple the luminance and the chromatic part. The 2-dimensional $C_b - C_r$ histogram was used as the base colour descriptor. Suitable colour statistics were extracted based on the footprint of the $C_b - C_r$ histogram, such as the Sarle's bi-modality index,

position, size and skew of the footprint. The breed-wise separability of the texture and colour feature vectors have been verified by projecting the feature vectors onto $\mathbb{R}^2$ space using t-SNE mapping technique.

- **Hierarchical Classification Scheme:** On the classification front, two different modalities have been proposed based on the distribution of the feature vectors in feature space. It was observed that because of some similarity in the colour and texture profile between different breeds, there was overlap between the different breeds in feature space. Also this overlap was found to be feature-type dependent. Thus the feature-types which provide optimal separation for different breed pairs were found to be different. This motivated the design of a multi-stage classification scheme in which each breed is siphoned out from the rest in a feature space where it is optimally separated from others and this siphoning takes place in a hierarchical fashion. A metric for computing the cluster distance between breed pairs as a function of three feature types: Colour(C), Texture(T) and combination of Colour and Texture features($C \cup T$) have been proposed. A cluster distance table has been created for the $\binom{4}{2} = 6$ breed pairs corresponding to the three feature types C, T and $C \cup T$. This table is fed as input to the hierarchical classification scheme which decides the order in which the breeds are to be siphoned out and the feature space to used at each stage of classification. An SVM classifier was used at each stage of classification to separate the two classes. The proposed hierarchical classification algorithm is found to outperform other similar classification methods in literature like the Phylogenetic Tree which uses the hierarchical clustering algorithm and Decision Trees. Classification accuracies of 86.45% for Duroc, 93.02% for Ghungroo, 86.91% for Hampshire and 98.54% for Yorkshire have been obtained with the proposed classification modality.

- **Multi-class Classification using Outlier Detection Frame:** While excellent classification accuracies of around 93% and 99% have been obtained for Ghungroo and Yorkshire respectively with the hierarchical classification scheme, the accuracies for Duroc and Hampshire need further improvement. Limited training data along with the inter-

class similarity between the breeds and intra-class variability within a breed places severe restriction on the performance of SVM classifier which is used at each stage of the hierarchical classification scheme. This gives rise to the requirement of another classification modality which can perform better under these constraints. In order to make inference on a test feature vector regarding its breed label when it lies in the overlapping region of two or more classes, it was found that the use of a generative classifier model is more appropriate instead of a discriminative one. The likelihood scores assigned to a test feature vector using the representation learnt for each training class is used to make inference on the breed label of the test feature vector. It was found that in order to bind the training data from a particular class via some form of an association map, representation learnt using spanning tree models are the most appropriate. Based on spanning tree models learnt for each training class, outlier scores are generated for the test feature vector corresponding to each training class. The test feature vector is assigned the class label for which it has the lowest outlier score. It was shown that based on the MaxST representation learnt for each training class, the test feature vector could be assigned an outlier score which is consistent with its proximity to the training class. However, the outlier score generated by such a representation based on MaxST is insensitive to the location of test feature vector when it falls within the training cluster. A test feature vector which lies deeper within a training cluster or is surrounded by more number of training feature vectors is a less probable outlier for the training class. With this consideration, two different methods have been proposed to take into consideration the positional arrangement of the test feature vector with respect to the feature vectors of the training cluster. One method is based on MinST representation learnt for the training class and the other method involved calculating the outlier score depending upon how deep the test feature vector is embedded within the training cluster in $\mathbb{R}^2$ space after projecting the feature vectors from the higher dimensional feature space using Random Projection Matrices(RPM). Although both the methods lead to the same classification accuracy, but the method based on MinST representation is computationally much more efficient. Excellent classification

accuracies of around 93% have been obtained for both Duroc and Hampshire; classification accuracy of around 97% have been obtained for Ghungroo, while Yorkshire registered a classification accuracy of 100%. The nearly 7% jump in the classification accuracies of Duroc and Hampshire highlights the superior performance of this multi-class classification framework over the hierarchical classification method.

7.2 Future Research Directions

There is a scope of improvement in the current work as well as there are several other fronts related to the problem of pig breed identification which remain open. These are enumerated below:

- In most practical situations, a significant fraction of the on-field captured muzzle image is occupied by the background region. This requires that the adaptive ball fitting segmentation algorithm discussed in Chapter 2, be fed with a manually cropped muzzle image in order to reduce the background interference to an acceptable level. Once the cropped muzzle image has been obtained, the subsequent steps towards breed identification, including the adaptive ball-fitting segmentation algorithm, feature extraction and classification needs no manual intervention. The generation of a manually cropped muzzle image from the on-field captured image acts as a hindrance towards the development of an automatic breed identification system for pigs. To develop a fully automated system towards breed identification, it becomes essential to design segmentation technique which can do the necessary pre-processing of the on-field captured muzzle image before it can be fed to the adaptive ball fitting algorithm discussed in Chapter 2.

  A possible scheme for reducing the background interference can start with capturing a series of snapshots in the form of a video of the pig. The different frames in the video can be analysed based on an algorithm to detect and select good quality full frontal muzzle images free from blur. The full frontal muzzle image is extremely necessary if one wants to segregate the breeds based on shape parameters. Also blurred images corrupt the

textural details heavily by introducing noise. The set of selected images can be then be fed to a suitable segmentation algorithm which first isolates the full body of the pig from the background based on body colour and/or shape and subsequently extracts out the muzzle region from the body. Thus the main steps in the algorithm can be enumerated as:

- Detection of frames with frontal muzzle image. along with their blur/sharpness assessment.
- Full body detection of pig followed by face segmentation based on a suitable combination of shape, colour or any other descriptor from these frames.
- Blur/sharpness assessment of these frames, so that the muzzle image with the highest sharpness metric can be selected for further processing.

The detection of frames with full frontal muzzle images is a challenging task since there is no referential prior information available regarding either the background or the nature of the breed. If deep learning techniques are to be applied, sufficient labelled data repository is to be constructed related to full frontal muzzle images for the four breeds. The segmentation of the pig body from the background can be facilitated by motion parameters if the initial video of the pig movement is captured with a still background. Blur/sharpness assessment is another task which is to be customized in the context of the muzzle images of the breeds. These pre-processing steps can be taken up as future work.

- As has been mentioned in Chapter 1 and also discussed in Chapter 2, it is extremely difficult to extract out the muzzle contour precisely. A precise extraction of the muzzle contour is expected to boost the classification accuracies further. If the muzzle contour can be extracted precisely the associated shape descriptor can be used for breed classification. As has been observed several times during the course of discussion in the thesis that Duroc and Hampshire are the two most difficult breeds to be separated out. The Duroc muzzle contour has two notches in the upper half area of the muzzle, and this a distinguishing feature of Duroc as compared to the other breeds. Thus if muzzle contour

can be detected precisely, then Duroc-Ghungroo and Duroc-Hampshire confusion can be reduced significantly resulting in an increase in the classification accuracy for each of these breeds. Hence, the muzzle contour detection which is a really challenging task provides a definite scope for future research.

- The quality of the captured muzzle image plays a key role towards the development of effective colour and texture descriptors. The texture descriptors in particular are more affected by the quality of the muzzle image. Unlike a human subject, an animal does not cooperate while capturing its image. Since a pig does not tend to remain stable while taking a snapshot of its muzzle, hence the captured muzzle image suffers from either motion blur or out-of-focus blur. A blurred muzzle image when presented to the feature extraction and classification algorithm, is very likely to be processed improperly, resulting in wrong inference about the breed with which the muzzle image is associated. Thus, the design of a blur assessment algorithm customized to quantify the sharpness of the muzzle image will prove to be extremely useful. Depending on the sharpness index generated by the algorithm, the person in-charge of capturing the snapshots can take a decision on whether the image of the muzzle needs to be re-captured or not so that it can be reliably processed by the feature extraction and classification algorithm.

- In an attempt to improve the classification accuracies further, i.e. beyond 93% for Duroc and Hampshire and 97% for Ghungroo, deep learning approaches can also be explored. The principal deterrent to the application of deep learning approaches is the non-availability of a large muzzle database. The standard datasets available in literature like MNIST, Fashion MNIST, ImageNet, CIFAR-10 etc. for image classification purposes using deep neural networks consist of at least few tens of thousands of images. In contrast the size of pig muzzle database available for the breed classification problem is very small. Thus, data augmentation needs to be done to artificially increase the size of the muzzle database. In addition to the standard data augmentation techniques, wherein augmented images are generated as rotated, translated or scaled version of the available images in the

original dataset, the use of Generative Adversarial Networks(GAN) can also be explored for generating augmented muzzle database. Thus, the search for suitable data augmentation technique along with the associated deep neural network architecture which can provide classification accuracy close to 100% for all the four breeds can be taken up as a future research work.

- Other biometrics can also be investigated as potential breed identifiers in pigs. For example, the body colour of the pigs can be a powerful descriptor. Ghungroo pigs are completely black, Duroc pigs have a brownish hue; Hampshire pigs are also black but with a prominent pink coloured belt located around its neck, Yorkshire pigs have a pinkish white colour. Apart from this the facial profile, ear shape, venation patterns in the ear etc. can also be investigated as potential breed identifiers.

- The current work described in this thesis is concerned with associating a muzzle image of a pig with one of the four breed types viz. Duroc, Ghungroo, Hampshire and Yorkshire, which are of high commercial significance in India. However, there are many more variety of domestic breeds available worldwide. It is highly impractical both from the point of view of market utility as well as design of discriminative features to consider the idea of identifying the breeds based on their muzzle images. However, it is quite possible that given a muzzle image, it actually does not belong to any of the four breeds. Hence, it becomes important to design robust visual descriptors and learn stable representations from them which can bind these four breeds together as a single entity, so that a muzzle image from a breed which falls outside these four categories can be recognized.

Bibliography

[1] R. C. Gonzalez, *Digital image processing*. Pearson education india, 2009.

[2] D. M. Tax and R. P. Duin, "Support vector data description," *Machine learning*, vol. 54, no. 1, pp. 45–66, 2004.

[3] P. Chetia, "High milk producing indian cattle breed: These 4 indian breed can give milk up to 80 liters," 2020, https://krishijagran.com/animal-husbandry/high-milk-producing-indian-cattle-breed-these-4-indian-breed-can-give-milk-up-to-80-liters/.

[4] S. Kadam, "Imported breeds of pig used in india," https://www.notesonzoology.com/india/pig-farming/imported-breeds-of-pig-used-in-india/1296.

[5] S. Chakraborty, K. Karthik, and S. Banik, "Investigation on the muzzle of a pig as a biometric for breed identification," in *Proceedings of 3rd International Conference on Computer Vision and Image Processing*. Springer, 2020, pp. 71–83.

[6] C. Cai and J. Li, "Cattle face recognition using local binary pattern descriptor," in *2013 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*. IEEE, 2013, pp. 1–4.

[7] B. F. B. Sampaio, C. E. S. N. Zúccari, M. Y. M. Shiroma, B. R. Bertozzo, E. C. R. Leonel, R. d. S. Surjus, M. M. M. Gomes, and E. V. d. Costa e Silva, "Biometric hoof evaluation of athletic horses of show jumping, barrel, long rope and polo modalities," *Revista Brasileira de Saúde e Produção Animal*, vol. 14, pp. 448–459, 2013.

[8] K. Karthik, S. Chakraborty, and S. Banik, "Muzzle analysis for biometric identification of pigs," in *2017 Ninth International Conference on Advances in Pattern Recognition (ICAPR)*. IEEE, 2017, pp. 1–6.

[9] Z. Weixing and Z. Jin, "Identification of abnormal gait of pigs based on video analysis," in *2010 Third International Symposium on Knowledge Acquisition and Modeling*. IEEE, 2010, pp. 394–397.

[10] R. Bardeli, "Similarity search in animal sound databases," *IEEE Transactions on Multimedia*, vol. 11, no. 1, pp. 68–76, 2008.

[11] R. Curtis, L. Viel, S. McGuirk, O. Radostits, and F. Harris, "Lung sounds in cattle, horses, sheep and goats," *The Canadian Veterinary Journal*, vol. 27, no. 4, p. 170, 1986.

[12] L. Lefèvre, E. Courtiol, S. Garcia, M. Thévenet, B. Messaoudi, and N. Buonviso, "Significance of sniffing pattern during the acquisition of an olfactory discrimination task," *Behavioural brain research*, vol. 312, pp. 341–354, 2016.

[13] H. B. Kim, K. Borewicz, B. A. White, R. S. Singer, S. Sreevatsan, Z. J. Tu, and R. E. Isaacson, "Longitudinal investigation of the age-related bacterial diversity in the feces of commercial pigs," *Veterinary microbiology*, vol. 153, no. 1-2, pp. 124–133, 2011.

[14] C. L. Collins, J. R. Pluske, R. S. Morrison, T. N. McDonald, R. J. Smits, D. J. Henman, I. Stensland, and F. R. Dunshea, "Post-weaning and whole-of-life performance of pigs is determined by live weight at weaning and the complexity of the diet fed after weaning," *Animal Nutrition*, vol. 3, no. 4, pp. 372–379, 2017.

[15] Y. Lu, X. He, Y. Wen, and P. S. Wang, "A new cow identification system based on iris analysis and recognition," *International journal of biometrics*, vol. 6, no. 1, pp. 18–32, 2014.

[16] S. Dan, S. Das, S. Mustafi, K. Roy, K. Mukherjee, S. N. Mandal, S. Banik, and S. Naskar, "Individual identification of black pig through ear images using support vector machine," in *2021 4th International Conference on Recent Trends in Computer Science and Technology (ICRTCST)*, 2022, pp. 169–174.

[17] K. Karthik, S. Chakraborty, and S. Banik, "Muzzle analysis for biometric identification of pigs," in *2017 Ninth International Conference on Advances in Pattern Recognition (ICAPR)*, 2017, pp. 1–6.

[18] W. Kusakunniran, A. Wiratsudakul, U. Chuachan, S. Kanchanapreechakorn, T. Imaromkul, N. Suksriupatham, and K. Thongkanchorn, "Biometric for cattle identification using muzzle patterns," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 34, no. 12, p. 2056007, 2020.

[19] PorkmoneyBlog, "Pig snouts: All you need to know about their importance and uses," 2019, https://www.porkmoney.com/blog/2019/07/05/pig-snouts-all-you-need-to-know-about-their-importance-and-uses/.

[20] W. Xun, L. Shi, H. Zhou, G. Hou, and T. Cao, "Effect of weaning age on intestinal mucosal morphology, permeability, gene expression of tight junction proteins, cytokines and secretory iga in wuzhishan mini piglets," *Italian Journal of Animal Science*, vol. 17, no. 4, pp. 976–983, 2018.

[21] L. Armi and S. Fekri-Ershad, "Texture image analysis and texture classification methods-a review," *arXiv preprint arXiv:1904.06554*, 2019.

[22] R. Pfister, K. A. Schwarz, M. Janczyk, R. Dale, and J. Freeman, "Good things peak in pairs: a note on the bimodality coefficient," *Frontiers in psychology*, vol. 4, p. 700, 2013.

[23] A. K. Jain and F. Farrokhnia, "Unsupervised texture segmentation using gabor filters," *Pattern recognition*, vol. 24, no. 12, pp. 1167–1186, 1991.

[24] L. L. Cavalli-Sforza and A. W. Edwards, "Phylogenetic analysis. models and estimation procedures," *American journal of human genetics*, vol. 19, no. 3 Pt 1, p. 233, 1967.

[25] R. Henson and L. Cetto, "The matlab bioinformatics toolbox. encyclopedia of genetics, genomics, proteomics and bioinformatics," 2005.

[26] A. S. R. Kumar, "Muzzle image analysis for individual pig and breed identification," Master's thesis, Indian Institute of Technology,Guwahati, India, 2016.

[27] E. Iqbal, A. Niaz, A. A. Memon, U. Asim, and K. N. Choi, "Saliency-driven active contour model for image segmentation," *IEEE Access*, vol. 8, pp. 208 978–208 991, 2020.

[28] A. Munir, S. Soomro, M. T. Shahid, T. A. Soomro, and K. N. Choi, "Hybrid active contours driven by edge and region fitting energies based on p-laplace equation," *IEEE Access*, vol. 7, pp. 135 399–135 412, 2019.

[29] C. Li, C. Xu, C. Gui, and M. D. Fox, "Distance regularized level set evolution and its application to image segmentation," *IEEE transactions on image processing*, vol. 19, no. 12, pp. 3243–3254, 2010.

[30] T. F. Chan and L. A. Vese, "Active contours without edges," *IEEE Transactions on image processing*, vol. 10, no. 2, pp. 266–277, 2001.

[31] L. Wang, J. Zhu, M. Sheng, A. Cribb, S. Zhu, and J. Pu, "Simultaneous segmentation and bias field estimation using local fitted images," *Pattern recognition*, vol. 74, pp. 145–155, 2018.

[32] X.-F. Wang, D.-S. Huang, and H. Xu, "An efficient local chan–vese model for image segmentation," *Pattern Recognition*, vol. 43, no. 3, pp. 603–618, 2010.

[33] Q. Cai, H. Liu, Y. Qian, S. Zhou, X. Duan, and Y.-H. Yang, "Saliency-guided level set model for automatic object segmentation," *Pattern Recognition*, vol. 93, pp. 147–163, 2019.

[34] X.-H. Zhi and H.-B. Shen, "Saliency driven region-edge-based top down level set evolution reveals the asynchronous focus in image segmentation," *Pattern Recognition*, vol. 80, pp. 241–255, 2018.

[35] Y. Zhang, W. Li, L. Zhang, X. Ning, L. Sun, and Y. Lu, "Adaptive learning gabor filter for finger-vein recognition," *IEEE Access*, vol. 7, pp. 159 821–159 830, 2019.

[36] X. Wang, W. Du, F. Guo, and S. Hu, "Leaf recognition based on elliptical half gabor and maximum gap local line direction pattern," *IEEE Access*, vol. 8, pp. 39 175–39 183, 2020.

[37] R. De Maesschalck, D. Jouan-Rimbaud, and D. L. Massart, "The mahalanobis distance," *Chemometrics and intelligent laboratory systems*, vol. 50, no. 1, pp. 1–18, 2000.

[38] L. v. d. Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of machine learning research*, vol. 9, no. Nov, pp. 2579–2605, 2008.

[39] P. Sebastian, Y. V. Voon, and R. Comley, "The effect of colour space on tracking robustness," in *2008 3rd IEEE Conference on Industrial Electronics and Applications*. IEEE, 2008, pp. 2512–2516.

[40] T. R. Knapp, "Bimodality revisited," *Journal of Modern Applied Statistical Methods*, vol. 6, no. 1, p. 3, 2007.

[41] W.-H. Li, "Simple method for constructing phylogenetic trees from distance matrices," *Proceedings of the National Academy of Sciences*, vol. 78, no. 2, pp. 1085–1089, 1981.

[42] L. Kannan and W. C. Wheeler, "Maximum parsimony on phylogenetic networks," *Algorithms for Molecular Biology*, vol. 7, no. 1, p. 9, 2012.

[43] NCBI, "Maximum likelihood," 2004, https://www.ncbi.nlm.nih.gov/Class/NAWBIS/Modules/Phylogenetics/phylo15.html.

[44] L. Kaufman and P. J. Rousseeuw, *Finding groups in data: an introduction to cluster analysis*. John Wiley & Sons, 2009, vol. 344.

[45] L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen, *Classification and regression trees*. CRC press, 1984.

[46] W.-Y. Loh and Y.-S. Shih, "Split selection methods for classification trees," *Statistica sinica*, pp. 815–840, 1997.

[47] J. R. Quinlan, "Improved use of continuous attributes in c4. 5," *Journal of artificial intelligence research*, vol. 4, pp. 77–90, 1996.

[48] F. Wang, Q. Wang, F. Nie, Z. Li, W. Yu, and F. Ren, "A linear multivariate binary decision tree classifier based on k-means splitting," *Pattern Recognition*, vol. 107, p. 107521, 2020.

[49] B. Chandra, R. Kothari, and P. Paul, "A new node splitting measure for decision tree construction," *Pattern Recognition*, vol. 43, no. 8, pp. 2725–2731, 2010.

[50] R. O. Duda, P. E. Hart *et al.*, *Pattern classification*. John Wiley & Sons, 2006.

[51] W. Menke, "Chapter 2 - some comments on probability theory," in *Geophysical Data Analysis (Fourth Edition)*, fourth edition ed., W. Menke, Ed. Academic Press, 2018, pp. 17–37. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780128135556000022

[52] I. J. Myung, "Tutorial on maximum likelihood estimation," *Journal of mathematical Psychology*, vol. 47, no. 1, pp. 90–100, 2003.

[53] S. Chakraborty, K. Karthik, and S. Banik, "Graph synthesis for pig breed classification from muzzle images," *IEEE Access*, vol. 9, pp. 127 240–127 258, 2021.

[54] I. C. of Agricultural Research, "Ghungroo pig: A potential strain of indigenous pig for the rural farmers," http://https://icar.org.in/node/8078.

[55] S. Kadam, "Imported breeds of pig used in india," https://www.notesonzoology.com/india/pig-farming/imported-breeds-of-pig-used-in-india/1296.

[56] M. Alam, M. D. Samad, L. Vidyaratne, A. Glandon, and K. M. Iftekharuddin, "Survey on deep neural networks in speech and vision systems," *Neurocomputing*, vol. 417, pp. 302–321, 2020.

[57] P. Chong, N.-M. Cheung, Y. Elovici, and A. Binder, "Toward scalable and unified example-based explanation and outlier detection," *IEEE Transactions on Image Processing*, vol. 31, pp. 525–540, 2021.

[58] A. Creswell, T. White, V. Dumoulin, K. Arulkumaran, B. Sengupta, and A. A. Bharath, "Generative adversarial networks: An overview," *IEEE Signal Processing Magazine*, vol. 35, no. 1, pp. 53–65, 2018.

[59] S. R. Gunn *et al.*, "Support vector machines for classification and regression," *ISIS technical report*, vol. 14, no. 1, pp. 5–16, 1998.

[60] W. Tang, K. Mao, L. O. Mak, and G. W. Ng, "Classification for overlapping classes using optimized overlapping region detection and soft decision," in *2010 13th International Conference on Information Fusion*. IEEE, 2010, pp. 1–8.

[61] J. E. Van Engelen and H. H. Hoos, "A survey on semi-supervised learning," *Machine Learning*, vol. 109, no. 2, pp. 373–440, 2020.

[62] S. J. Peter, "Minimum spanning tree based clustering for outlier detection," *Journal of Discrete Mathematical Sciences and Cryptography*, vol. 14, no. 2, pp. 149–166, 2011.

[63] S. Abghari, V. Boeva, N. Lavesson, H. Grahn, S. Ickin, and J. Gustafsson, "A minimum spanning tree clustering approach for outlier detection in event sequences," in *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*. IEEE, 2018, pp. 1123–1130.

[64] P. Juszczak, D. M. Tax, E. Pe, R. P. Duin *et al.*, "Minimum spanning tree based one-class classifier," *Neurocomputing*, vol. 72, no. 7-9, pp. 1859–1869, 2009.

[65] R. La Grassa, I. Gallo, A. Calefati, and D. Ognibene, "Binary classification using pairs of minimum spanning trees or n-ary trees," in *International Conference on Computer Analysis of Images and Patterns*. Springer, 2019, pp. 365–376.

[66] R. La Grassa, I. Gallo, and N. Landro, "Ocmst: One-class novelty detection using convolutional neural network and minimum spanning trees," *Pattern Recognition Letters*, 2021.

[67] F. Gavril, "Generating the maximum spanning trees of a weighted graph," *Journal of Algorithms*, vol. 8, no. 4, pp. 592–597, 1987.

[68] D. Kwon and H. Ju, "A media distribution tree construction method using a maximum spanning tree for mobile smart devices," in *2015 17th Asia-Pacific Network Operations and Management Symposium (APNOMS)*. IEEE, 2015, pp. 227–231.

[69] H.-Y. Ha, S.-C. Chen, and M. Chen, "Fc-mst: Feature correlation maximum spanning tree for multimedia concept classification," in *Proceedings of the 2015 IEEE 9th International Conference on Semantic Computing (IEEE ICSC 2015)*. IEEE, 2015, pp. 276–283.

[70] Y. Huang, X. Li, and W. Ai, "Lossless compression of modis data based on the maximum spanning tree and 3d context prediction," in *2011 Fourth International Conference on Intelligent Computation Technology and Automation*, vol. 2. IEEE, 2011, pp. 596–599.

[71] X. Wang, X. Wen, K. Ma, and D. Zhang, "A multilayer maximum spanning tree kernel for brain networks," in *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*. IEEE, 2021, pp. 1582–1585.

[72] J. Chang, J. Luo, J. Z. Huang, S. Feng, and J. Fan, "Minimum spanning tree based classification model for massive data with mapreduce implementation," in *2010 IEEE International Conference on Data Mining Workshops*. IEEE, 2010, pp. 129–137.

[73] X. Wang, X. L. Wang, and D. M. Wilkes, "A minimum spanning tree-inspired clustering-based outlier detection technique," in *Industrial Conference on Data Mining*. Springer, 2012, pp. 209–223.

[74] M. A. Haque and H. Mineno, "Contextual outlier detection in sensor data using minimum spanning tree based clustering," in *2018 International Conference on Computer, Communication, Chemical, Material and Electronic Engineering (IC4ME2)*. IEEE, 2018, pp. 1–4.

[75] R. C. Prim, "Shortest connection networks and some generalizations," *The Bell System Technical Journal*, vol. 36, no. 6, pp. 1389–1401, 1957.

[76] F. Busato and N. Bombieri, "Graph algorithms on gpus," *Adv. GPU Res. Pract*, pp. 163–198, 2017.

[77] W. Mathworld, "Triangle interior," https://mathworld.wolfram.com/TriangleInterior.html.

[78] E. Bingham and H. Mannila, "Random projection in dimensionality reduction: applications to image and text data," in *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, 2001, pp. 245–250.

[79] B. J. William and J. Lindenstrauss, "Extensions of lipschitz mapping into hilbert space," *Contemporary mathematics*, vol. 26, no. 189-206, pp. 323–341, 1984.

[80] B. Tang and H. He, "A local density-based approach for outlier detection," *Neurocomputing*, vol. 241, pp. 171–180, 2017.

[81] V. Hautamaki, I. Karkkainen, and P. Franti, "Outlier detection using k-nearest neighbour graph," in *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.*, vol. 3. IEEE, 2004, pp. 430–433.

[82] H. Huang, K. Mehrotra, and C. K. Mohan, "Rank-based outlier detection," *Journal of Statistical Computation and Simulation*, vol. 83, no. 3, pp. 518–531, 2013.

[83] S. Ramaswamy, R. Rastogi, and K. Shim, "Efficient algorithms for mining outliers from large data sets," in *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, 2000, pp. 427–438.

[84] J. Tang, Z. Chen, A. W.-C. Fu, and D. W. Cheung, "Enhancing effectiveness of outlier detections for low density patterns," in *Pacific-Asia conference on knowledge discovery and data mining*. Springer, 2002, pp. 535–548.

[85] J. Navarro, I. M. de Diego, R. R. Fernández, and J. M. Moguerza, "Triangle-based outlier detection," *Pattern Recognition Letters*, vol. 156, pp. 152–159, 2022.

[86] A. Daneshpazhouh and A. Sami, "Entropy-based outlier detection using semi-supervised approach with few positive examples," *Pattern Recognition Letters*, vol. 49, pp. 77–84, 2014.

[87] M. Riahi-Madvar, A. A. Azirani, B. Nasersharif, and B. Raahemi, "A new density-based subspace selection method using mutual information for high dimensional outlier detection," *Knowledge-Based Systems*, vol. 216, p. 106733, 2021.

[88] F. Keller, E. Muller, and K. Bohm, "Hics: High contrast subspaces for density-based outlier ranking," in *2012 IEEE 28th international conference on data engineering*. IEEE, 2012, pp. 1037–1048.

[89] M. Riahi-Madvar, B. Nasersharif, and A. A. Azirani, "Subspace outlier detection in high dimensional data using ensemble of pca-based subspaces," in *2021 26th International Computer Conference, Computer Society of Iran (CSICC)*. IEEE, 2021, pp. 1–5.

[90] C. Blake, "Uci repository of machine learning databases," *http://www. ics. uci. edu/˜ mlearn/MLRepository. html*, 1998.

[91] A. P. Bradley, "The use of the area under the roc curve in the evaluation of machine learning algorithms," *Pattern recognition*, vol. 30, no. 7, pp. 1145–1159, 1997.

[92] M. Verleysen and D. François, "The curse of dimensionality in data mining and time series prediction," in *International work-conference on artificial neural networks*. Springer, 2005, pp. 758–770.

[93] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander, "Lof: identifying density-based local outliers," in *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, 2000, pp. 93–104.

List of Publications

Journal Publications

- Published:

  1. Chakraborty, S., Karthik, K. and Banik, S., 2021. Graph Synthesis for Pig Breed Classification From Muzzle Images. IEEE Access, 9, pp.127240-127258.

- Journals under preparation

  1. Chakraborty, S., and Karthik, K., Outlier Detection Framework based on Spanning Trees.

Conference Publications

1. Karthik, K., Chakraborty, S. and Banik, S., 2017, December. Muzzle analysis for biometric identification of pigs. In 2017 Ninth International Conference on Advances in Pattern Recognition (ICAPR) (pp. 1-6). IEEE.

2. Chakraborty, S., Karthik, K. and Banik, S., 2020. Investigation on the muzzle of a pig as a biometric for breed identification. In Proceedings of 3rd International Conference on Computer Vision and Image Processing (pp. 71-83). Springer, Singapore.