ALARM CALL ANALYSIS OF FARM PIGS USING DEEP LEARNING AND NOISE FILTERING METHODS

KAUSHIK DEY

THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE IN DATA SCIENCE & SOCIETY
AT THE SCHOOL OF HUMANITIES AND DIGITAL SCIENCES
OF TILBURG UNIVERSITY

STUDENT NUMBER

2060967

COMMITTEE

dr. Dan Stowell
prof. dr. Henry Brighton

LOCATION

Tilburg University
School of Humanities and Digital Sciences
Department of Cognitive Science &
Artificial Intelligence
Tilburg, The Netherlands

DATE

June 24, 2022

WORD COUNT

8320

ACKNOWLEDGMENTS

Thank you for taking your time to read my thesis which focuses on accurately classifying the alarm calls of pigs using Convolutional Neural Networks (CNN) and making performance comparisons between the models. Firstly, I would like to thank my supervisor, Dr. Dan Stowell, for guiding me throughout this Thesis journey. He has been an amazing supervisor and I found his feedback and support very helpful throughout my thesis. Also, I sincerely thank Pim van Gennip for providing me with the data and valuable guidance. Furthermore, I would like to thank my mother, family, and childhood friends for being strong pillars during this time. Finally, I hope that in reading this thesis you will learn as much as I did in writing and experimenting, and more importantly, that you will enjoy reading it!

CONTENTS

ALARM CALL ANALYSIS OF FARM PIGS USING DEEP LEARNING AND NOISE FILTERING METHODS

KAUSHIK DEY

Abstract

As we know that global meat consumption is increasing every year and in order to meet the demand, we have to think of the welfare of the farm animals as well. There are various information communications technologies that have aided the development of precision livestock farming (PLF), through which the livestock industry is pursuing welfare breeding (which improves livestock environments) and the production of quality livestock products (Meen et al., 2015; Nasirahmadi, Edwards, & Sturm, 2017). But the current progress in the research is not enough to meet the demand of the ever-increasing billion population and most of the research work has been in the field of detecting the emotional state of the animals or detecting the animal sounds in general. It is, for this reason, this thesis is going deep into the problem of detecting the physical health of the farm pigs and helping the research of animal welfare. The vital part of such PLF is the technology to accurately monitor the current conditions of the livestock. And animal 'alarm call' could also be a key indicator of animal health issues. The alarm calls in nonhuman animals are such vocal sounds that are typically produced in response to a perceived danger in their surroundings as well as animals phonate alarm calls to exhibit their own physical health-related illness/distress (Price et al., 2015). And these calls are acoustically distinct and have a specific set of acoustic features which are different from normal vocal sounds (Chan, Cloutier, & Newberry, 2011). Alarm calls such as coughs and sneezes could be detected in pig farms to check health-related problems in pigs and the farm owners can be alerted immediately in case any health-related anomaly is found. In this project, two machine-learning models have been trained to detect pig alarm calls such as coughing and sneezing as well as normal calls, using data from a real commercial product, and the performance of the algorithms has been compared against each other. The first deep learning model called the 'Efficient-Bo' model has achieved an accuracy of 94.29% and 89.89% and the second model 'ResNet-18' has obtained an accuracy of 88.45% and 77.61% with noisy and denoised data respectively.

1 DATA SOURCE AND CODE

This thesis is based on the collection of animal data from the local farms in the Netherlands. But the data collection policy and experimental design have complied with the rules and regulations of the TSHD Research Ethics and Data Committee for the requirements of the GDPR EU. The data that has been used for this thesis was provided by an external organization, which is the original owner of the data during and after the completion of this thesis. Moreover, this data is not publicly available to any other entity. Also, I do not hold any legal claim or authority over the data used in this thesis project.

2 INTRODUCTION

2.1 *Motivation*

Global meat consumption has increased with an exponential curve because of the population and rapid economic growth all around the world and in order to meet the ever-increasing demand for livestock products, the livestock industry has expanded and implemented dense breeding (Meen et al., 2015). Additionally, various information communications technologies have aided the development of precision livestock farming (PLF), through which the livestock industry is pursuing welfare breeding (which improves livestock rearing environments) and the production of quality livestock products (Nasirahmadi et al., 2017). Also, disease surveillance and prevention have become a critical and important challenge in current livestock production (Aarnink, Swierstra, Van den Berg, & Speelman, 1997). Early detection of disease is an effective way to prevent massive disease outbreaks and save consequent economic costs. Respiratory diseases are one of the most common and frequently-occurring diseases in pig husbandry (Greiner, Stahly, & Stabel, 2000). They generally slow down the feed conversion and growth rate of the pigs and end up making the pork extremely low quality (Pijpers et al., 1991). It also increases the farm management and treatment costs of the livestock and more likely increases the death rate of the pigs in serious health-related instances (Greiner et al., 2000). Coughing and other alarm calls are one of the first-hand symptoms of lung-related diseases. The alarm call detection shows us a way to build a Machine Learning and IoT-enabled intelligent system for detecting diseases early on (Van Hirtum et al., 2003). An intelligent alarm system for respiratory diseases can help farmers treat sick animals to prevent the disease from spreading on the farm while maintaining the livestock quality and improving animal welfare. That is what makes this project

worthwhile. To counter the issues faced by modern livestock farming, this project has progressed to detect the alarm calls of the pigs early on through technological advancements in Data Science. But the project had not been sucessful without the active cooperation from the company named Iconize BV. The audio data is collected by Iconize BV using the Healthy Climate Monitor equipment. Iconize BV, which was created in 2005, focuses on producing software and hardware with a focus on long-term sustainability and Healthy Climate Monitor is a device that uses IoT sensors to capture live photos and audio recordings and analyzes the data in real-time.

2.2 *Scientific and Societal Relevance*

This research project is the first attempt to distinguish sneezing from normal calls by using machine learning algorithms and comparing their performances against each other. Most of the literature on improving animal welfare has been focused on detecting animal vocalizations mainly for investigating possible statistical correlations between general changes in behavior as responses to different types of tests (Murphy, Nordquist, & van der Staay, 2014). Animal voices or sounds have been used to a large extent to conduct research on classifying different animal species (Chen, Huang, Chen, Chen, & Chien, 2015). But on the other hand, classification studies on the behaviors and statuses of specific animals have been rarely conducted (Sauvé, Beauplet, Hammill, & Charrier, 2015). So, it is for this reason, that the machine learning approach in this project brings relevance to the lack of enough technical research for this particular problem faced in modern livestock farming. Also, this project is equally motivated by a goal to bring significant values to our society and make the lives of the farm pigs better, and increase the economic growth of the Precision Livestock Farming(PLF) industry. The important aspect of such PLF is to implement a technology that can accurately monitor the current conditions of the livestock. This is where the research behind this project is playing a crucial role.

2.3 *Research Question(s)*

In the context of this research project, the following research questions have been formulated.

RQ1 *To what extent can a supervised Convolutional Neural Network (CNN) algorithm discriminate between normal calls and alarm calls such as coughing, and sneezing made by the pigs on a rural farm?*

A sub-question has also been framed, as such:

RQ2 *With how much accuracy can the algorithm in question classify these different calls correctly in the absence of background noise in the audio data?*

For differentiating the alarm calls from the normal calls, two convolutional neural networks (CNN) such as EfficientNet-Bo and ResNet-18 have been built. To tackle the issue of the background noise in the audio data and finally remove it, a noise-filtering method called 'Spectral Subtraction' has been applied to all the pig sound recordings.

3 RELATED WORK

The search for literature for this thesis project has been based on two important aspects - 1) Cattle alarm call and its relevance to Animal Welfare and 2) Alarm call analysis using machine learning approach. The first aspect focused on the animal call vocalization and its relevance to animal welfare. A handful of research has been conducted in the past to find the link between animal call sounds and their physical health states and how it is significantly relevant to animal welfare. Based on this theoretical evidence and findings, the other aspect focuses on the discussion of the development of the machine learning model to analyze the cattle vocal sounds, go over their findings, and their overall research limitation based on which this current research project has been extended and further progressed.

3.1 *Cattle alarm call and its relevance to Animal Welfare*

Vocal information acts as an important communication tool between animal groups or individuals that has a big advantage of effectively reaching a long range of distance. The research conducted by (Jung, Kim, Moon, Kim, et al., 2021) serves the purpose of conveying the information about health-related anomalies found through vocal analysis of laying hens and cattle. The voice of the vocalizing animal gives us information about the age, gender, sequence, and breeding status of the animals. The voice of cattle, therefore, carries vital information about the animal's extraordinary conditions, such as pain, estrus, separation from the calf, and hunger or thirst (Ikeda & Ishii, 2008; Riede, Tembrock, Herzel, & Brunnberg, 1997). When the cattle feel hungry or thirsty or are in pain (Yeon et al., 2006) or in stressful circumstances, the cattle mimic a voice to indicate meaning.

In order to meet the increasing global demand for livestock products, practices in livestock management have moved on to more intensive methods (Grandin, 2014). Even though output has increased to a large extent

(P. K. Thornton, 2010), it is still difficult for farm owners to carefully scrutinize and monitor every individual cattle. At the same time, consumers of the livestock products are asking for more transparency in the animal welfare, raising questions about the environmental impact, and asking for safety protocol which has to be taken for the animal products they purchase (Grandin, 2014; Moynagh, 2000; P. K. Thornton, 2010). Disease surveillance among cattle and prevention of it is very important as well as demanding which is a challenge that the current livestock production (Aarnink et al., 1997; Wathes, Jones, Kristensen, Jones, & Webster, 2002) industry is facing. This problem brings a dilemma for modern farm owners who are to find a proper balance between high production targets, ethical issues, sustainability for the environment, and safety requirements for the cattle animals. Precision Livestock Farming (PLF) has a target to address all these issues by making the monitoring of the farm livestock continuous and automated and enabling more appropriate and timely interventions (Vandermeulen et al., 2015). Because a failure in early detection of respiratory diseases and aggressive behaviors among weaned pigs due to the social conflict may result in substantial financial harm, these investigations are generally separated into identifying coughing sounds produced by diseases and screaming induced by stress (Cordeiro, Nääs, da Silva Leitão, de Almeida, & de Moura, 2018; Lee, Jin, Park, & Chung, 2016).

3.2 *Alarm call analysis using Machine Learning approach*

A convolutional neural network (CNN) is a deep learning system that stacks a two-dimensional data array, such as an image, using a series of two-dimensional filters. CNN has already been used in voice classification and has shown great accuracy in image classification (Khamparia et al., 2019; Nanni, Rigo, Lumini, & Brahnam, 2020; Oikarinen et al., 2019). For animal sound classification using CNNs, (Xie & Zhu, 2019) applies deep learning in classifying Australian bird sounds and reports a classification accuracy of more than 88%. Using a wireless acoustic sensor network, (Xu, Zhang, Yao, Xue, & Wei, 2020) suggest a multiple-view CNN architecture for classifying animal species. Despite advances in categorization technology that allows for real-time monitoring of voice data in the field of PLF, maintaining quality in the vocal capture system remains a challenge (J. C. Bishop, Falzon, Trotter, Kwan, & Meek, 2019; Green, Clark, Lomax, Favaro, & Reby, 2020; Mead-Hunter, Selvey, Rumchev, Netto, & Mullins, 2019). Also, (Chedad et al., 2001) presents a neural network-based approach for identifying cough noises from other sounds such as growls, metal clanging, and background noise. Several researchers have recently demonstrated successful noise reduction utilizing short-time Fourier transform (STFT) filters. Artificial

intelligence technology is, therefore, necessary to evaluate cattle's call by obtaining noise-removed voice information and evaluating it in a livestock facility in order to efficiently acquire sound and precisely determine this information (Chen et al., 2015).

Animal phonetics research has made use of audio and video recording technologies. (Meen et al., 2015) describe a feasible welfare monitoring system that uses audio and video recordings to track the vocalizations and behavior of Holstein Friesian cattle. All through the day, laying hens interact with one another through constant auditory input. Different chicken sounds have been employed in certain research to determine the health state of poultry (Huang, Wang, & Zhang, 2019; Sadeghi, Banakar, Khazaee, & Soleimani, 2015). Cattle phonate in a particular pattern when they are hungry or thirsty, in estrus, or under stress. Although several researchers (Meen et al., 2015; Xuan et al., 2016) have attempted to recover these acoustic sound characteristics and categorize each of the sounds, the classification accuracy is not significant enough, or rather the sound to be classified is simple.

The paper proposed by (Jung, Kim, Moon, Jhin, et al., 2021) implements a web-based and real-time cattle monitoring system that detect Coughing, Estrus Call, cattle Food-anticipation call, and Normal call of the cows. They have used a Sequential CNN model (see Figure 1) to classify the vocal sounds and STFT is used to remove the background noise. With the combination of the CNN and Mel-frequency cepstral coefficients (MFCCs), they have done the real-time cattle sound detection and an accuracy of 82% percent is obtained finally (Jung, Kim, Moon, Jhin, et al., 2021). Research

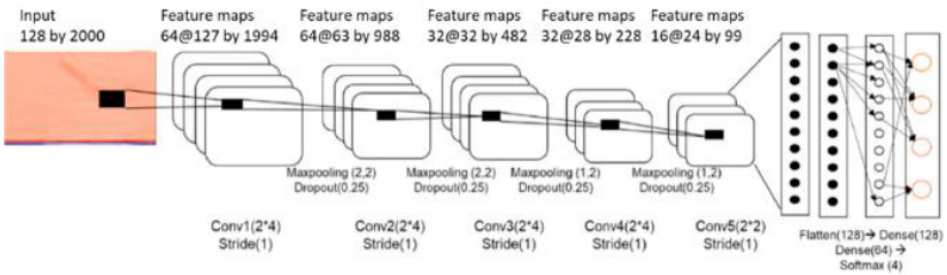

Figure 1: Sequential CNN for classifying cattle calls

advocated by (Yin, Tu, Shen, & Bao, 2021) provides a highly accurate pig cough recognition method for the respiratory disease alarm system. They implement a Classification algorithm based on fine-tuned AlexNet model (see Figure 2). With the help of CNN in the Image Recognition field, they convert sound signals into Spectrogram images for better accuracy. In the end, they have been able to obtain an accuracy of 96.8% and 95.4% for

cough and overall recognition, respectively (Yin et al., 2021). The research

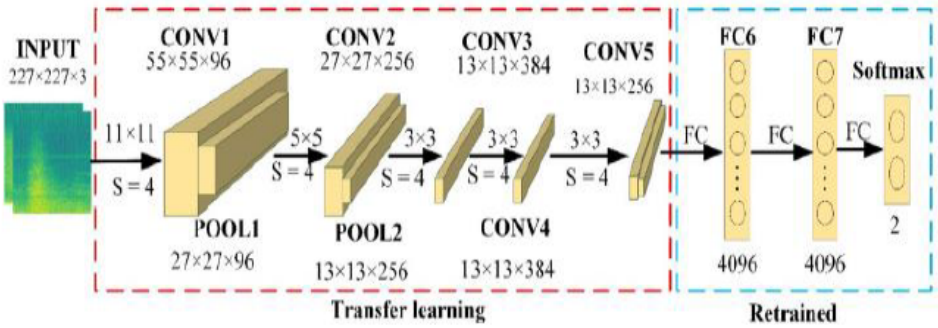

Figure 2: AlexNet CNN model

work proposed by (J. C. Bishop et al., 2019) have developed a multipurpose animal vocalization classification algorithm that is a combination of the machine learning model Support Vector Machine(SVM) and of MFCC and DWT for the feature extraction part as they find none of the algorithms currently available can easily be adapted and applied to different species. Then, they evaluate the performance of their algorithm on three different data sets targeting different animals such as dogs, sheep, and cattle where accuracies of 99.29%, 95.78%, and 99.67% are obtained for sheep, cattle, and dogs respectively (J. C. Bishop et al., 2019). Each dataset has been split into one-hour intervals by using Audacity audio editing software (Amiriparian et al., 2017), and those segmentations are afterward converted into the spectral domain using a method called Fast Fourier Transform (FFT). The FFT applies a window size of 1024 and uses a Hanning window function, which is selected for its balanced frequency resolution, sidelobe roll-off rate, and side lobe level reduction (Gaberson, 2006). Spectrogram images have been produced for each audio segmented data produced by the FFT and it is then visually inspected to make sure the overall level of vocal sound analysis has been performed correctly, using Sonic Visualiser software (Cannam, Landone, & Sandler, 2010). In the study done by (Hong et al., 2020), a system has been built to quickly detect various abnormalities (respiratory diseases, aggressive behaviors among pigs, etc). The project includes steps such as sound acquisition from sound sensors to signal-to-sound area detection and converting them into a spectrogram. Then, an 8-bit filtering clustering algorithm is applied to MnasNet (a lightweight deep learning model) and used to receive the spectrogram as input to detect and identify abnormal pig situations. The experiment obtains an F1 score of 0.947 (Hong et al., 2020). The research conducted by Jung, Kim, Moon, Kim, et al. (2021) serves the purpose of conveying the information about health-related anomalies found through vocal analysis of laying hens

and cattle. They develop two types of CNN algorithms; one is based on 2D

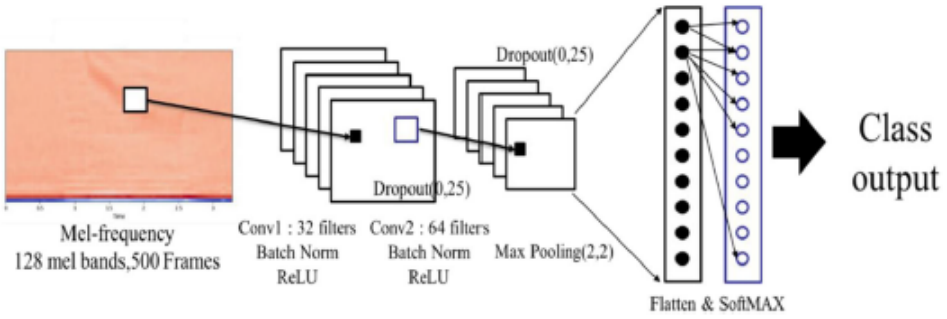

Figure 3: ConVnet CNN model

ConVnet and the second CNN is built on a 1D model with long short-term memory and both are compared against each other for evaluation purposes. The CNN model based on the 2D ConVnet (see Figure 3) performes better with an accuracy of 75.78 percent for laying hens and 91.02% for cattle (Jung, Kim, Moon, Kim, et al., 2021).

3.3 *Research gaps in literature*

After reviewing the works of literature, several literature gaps have been uncovered. For example, vocal sound differences between individual animals in a herd have not been included during the training of the model. Which particular animal in the whole herd makes which sound that cannot be detected. This particular problem can be tackled by installing a video imaging system at the farm. Also, not just coughing, future work can be done to distinguish between dry and wet cough or single and continuous cough, and so on because this information carries the information needed for medical treatment. In one case study, the algorithm does not perform well as the segmentation and extraction of the sound instances are manual; that can be improved by developing the automated segmentation component. Many research has also failed to combine audio data with video data collected from the sensors installed at the farms in order to find health-related anomalies in cattle. Interestingly, sometimes there has also been a shortage of enough training data for all classes of sounds/alarm calls. Given various research gaps, this thesis has focused on differentiating between normal calls and alarm calls taking enough training data for all classes of the sounds. Also, no previous research in the past has been conducted on detecting a particular alarm call called 'sneezing' in the pigs. Hence, this thesis work has also aimed for catching the 'sneezing' calls in the pigs.

4 METHOD

4.1 *Motivation behind using pre-trained CNN model over custom-made CNN model*

Modern Data Science image or time-series classification problems can be solved by using a pre-trained CNN model or a custom-made CNN in a traditional way. But (Erhan, Manzagol, Bengio, Bengio, & Vincent, 2009) suggests pre-training as an important mechanism for the parameters of the network. (Bengio, Lamblin, Popovici, & Larochelle, 2006) have also emphasized that pre-training is useful for initializing the network in such parameter space where optimization would be convenient, as well as a better local optimum of the training criterion, can be found. In the task of object recognition, one custom Convolutional Neural Network(CNN) and two pre-trained CNN models, namely AlexNet and GoogleNet are used and the experimental result reveals that custom CNN achieves an accuracy of 91.05% while AlexNet and GoogleNet obtains an accuracy score of about 99.65% (Zabir, Fazira, Ibrahim, & Sabri, 2018).

In the audio classification task, similar performances have been achieved. In one research where heartbeat sound is classified using EfficientNet-Bo and gets an accuracy of over 83% (Ul Haq et al., 2021). Similarly, in bird call recognition task where a pre-trained ResNet model is used and achieves an accuracy of more than 70% (Sankupellay & Konovalov, 2018). It is for this reason that this thesis has focused on using pre-trained CNN models, especially EfficientNet-Bo and ResNet-18, for categorizing pig alarm calls and normal calls and the performance is evaluated between the two models by using noisy audio data first and denoised audio data later on.

4.2 *Software and Packages*

The scripting language which has been used for the thesis is Python 3.7. Python libraries such as Pandas, Keras, PyDub, NumPy, Hydra, Librosa, Timm, and so on have been used for the development of the codes for the thesis. PyTorch has been used as a deep learning programming framework. Also, Anaconda Notebook, and Google Colab worked as the code implementation environments.

4.3 *Dataset*

4.3.1 *Source*

The raw audio data of the animal calls have been recorded in the LIVE settings at the rural farms in the Netherlands. The vocal sounds are recorded by microphones and several modern IoT devices, which have been installed at the farms in order to record the animal sounds. The company called Iconize BV collects the audio data using the Healthy Climate Monitor instrument. Iconize BV, founded in 2005, specializes in developing software and hardware products, keeping the main focus on sustainable development. The Healthy Climate Monitor is an instrument that works on live images and live audio recordings with the help of IoT sensors and promptly analyses the data.

Figure 4: Alarm Call recording of pigs at the rural farm

4.3.2 *Format and Size*

The audio dataset is the vocal sounds of the pigs recorded at various farms in the Netherlands. The dataset which has been used for training the Deep Learning models has been carefully labeled under the guidance of the animal doctors (veterinarians). The dataset also includes the background noise of the farm. There are six hundred and seventy annotated audio files in the FLAC file format. Through the Data Augmentation method, the

amount of data has reached 23180 annotated audio files. The total duration of the coughing, sneezing, and normal call recordings is three hundred eighty-six minutes and thirty-three seconds. The following table 1 shows the number of files in original format as well as in augmented format for each call.

Table 1: Alarm Call Dataset

Alarm Call	No. of original call	No. of augmented call	After augmentation
Coughing	15	7500	7515
Normal	629	7548	8177
Sneeze	26	7462	7488

4.4 *Preprocessing and Feature Extraction*

In this thesis, there are a few major pre-processing which have been performed on the data such as the conversion of the FLAC files to the WAV files with online conversion software for the convenience to work with the neural network model and Python scripting language. Secondly, as the labeling or annotating every single WAV audio file is time-consuming as well as economically expensive, the Data Augmentation method has been performed which basically increases the volume of the data by creating synthetic data out of the existing data. This results in enough data for training the neural network. Afterward, all the alarm call recordings have been denoised (removal of the noise from the audio files) in order to see how accurately the model predicts in case of the absence of background noise. Audio noise can be effectively suppressed via Spectral Subtraction. The Spectral Subtraction algorithm is a well-known technique for voice enhancement that was first described by Boll (Kim, Caire, & Molisch, 2015). To attain a higher signal-to-noise (SNR) ratio, the average signal spectra and average noise spectra are computed and subtracted (Lei, Lin, He, & Zuo, 2013). This noise-filtering method is called 'Spectral Subtraction' which has been applied to the noisy data.

4.4.1 *Data Augmentation*

Although the model design and hyperparameter tweaking are important aspects of creating a great model. Intuitively, one of the most typical issues in real data science problems is a shortage of data. Data augmentation, or the application of one or more deformations to a collection of annotated training samples to produce new, additional training data is an elegant solution to this problem Krizhevsky, Sutskever, and Hinton (2012). Data

augmentation facilitates the generation of synthetic data from existing data sets, enabling the model's generalization power to be increased (Simard, Steinkraus, Platt, et al., 2003). The deformations given to the labeled data do not modify the semantic meaning of the labels, which is an important idea in data augmentation (McFee, Humphrey, & Bello, 2015). Out of all the data augmentation methods in existence, only three audio data augmentation methods that have been applied to each audio file in this project are Noise Injection, Pitch Shifting, and Band Stop Filter. For audio data augmentation, the primary libraries which have been used are Librosa, Audiomentation, and PyDub.

With Noise Injection, there are three types of noise which are Gaussian noise, Salt and Pepper noise, and Speckle noise. In this project, Gaussian noise has been added keeping the minimum and maximum amplitude at 0.01 and 0.02. For Pitch shifting, this augmentation method changes the pitch of alarm call sound randomly using the Librosa function from Python. The maximum and minimum semitone has been kept at 8 and -8 respectively. Then, the Band-stop Filter is applied to the input audio file. This is also known as Notch Filter or Band Reject Filter. The maximum and minimum center frequency has been kept at 100 and 10000 respectively which has been decided based on the value of the sampling rate of audio files. The following figures show the Mel Spectrogram of the original audio (see Figure 5) and the Mel spectrogram of the same audio after all the three augmentation methods are applied to it (see Figure 6).

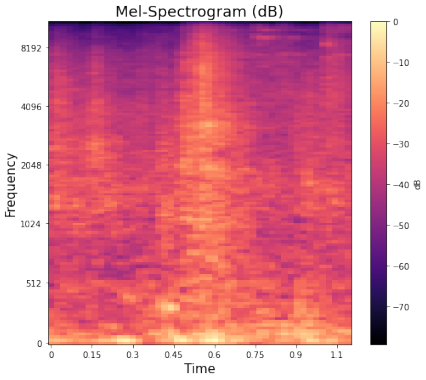

Figure 5: Mel Spectrogram of Raw Audio

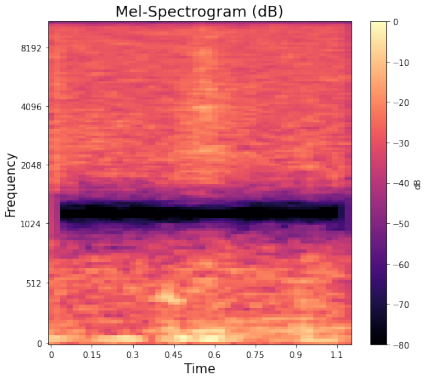

Figure 6: Mel Spectrogram of Augmented Audio

4.4.2 *Spectral Subtraction*

After the Data Augmentation process, the Spectral Subtraction method on the data to remove the noise from the audio data has been applied. The spectral subtraction approach works by subtracting the noise magni-

tude spectrum from the noisy speech spectrum (Biswas, Pal, Mandal, & Chakrabarti, 2014). The spectral subtraction approach also provides the following benefits: low computing complexity, great real-time performance, ease of implementation, and so on (Biswas et al., 2014).

Noise reduction cannot be done in the time domain; instead, it must be done in the frequency domain. The spectral subtraction noise removal method used in this thesis involves segmenting the noisy speech signal into half-overlapped time-domain data buffers multiplied by a Hanning window (discussed in the following paragraph) and then transforming the result into the frequency domain using the fast Fourier transform (FFT). After that, noise is eliminated by subtracting the average magnitude of the noise spectrum from the noisy speech spectrum and using half-wave rectification to zero out the negative values. Finally, the noise-reduced speech is rebuilt back into the time domain using the Inverse Fast Fourier Transform (IFFT) once the noise has been removed (Rabiner, 1978). The noisy voice data is decomposed into time windows known as Hanning time windows in this study. The Hanning time window is a bell-shaped curve that multiplies the noisy voice data from the half-overlapped data buffers. Outside the Hanning time frame, some of the noisy speech data is canceled out, while sections within are further assessed for processing. A Hanning time window's mathematical general formulation is as follows (Karam, Khazaal, Aglan, & Cole, 2014) :

$$W[n] = \begin{cases} 0.5 - 0.5 \times \cos\left(\frac{2 \times \pi}{L} n\right) \\ 0 \end{cases} \tag{1}$$

where L is the Hanning time window's length or the number of samples. The data in the Hanning time window is assessed for spectral computation, which requires applying the FFT technique to compute the discrete Fourier transform (DFT).

The following figures show the Mel Spectrogram of the original audio (see Figure 7) and the Mel Spectrogram of the same audio after the Spectral Subtraction method is applied to it (see Figure 8).

4.4.3 *Data Normalization and Mel Spectrograms*

The feature which is being extracted from the audio files is called the Mel Spectrogram. A spectrogram is a visual representation of the frequency content of a signal across time. The Mel scale is a linear scale for the human auditory system that is connected to Hertz using the following formula (B. Thornton, 2019):

$$m = 2595 \times \log_{10}(1 + f/700) \tag{2}$$

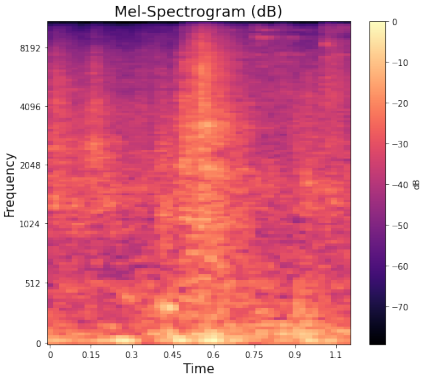

Figure 7: Mel Spectrogram of Raw Audio

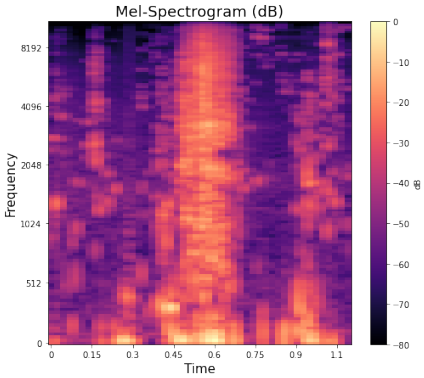

Figure 8: Mel Spectrogram of Denoised Audio

The Mel Spectrogram is used to provide the models with sound information similar to what a human would perceive. In the project, the raw audio waveforms are passed through filter banks to obtain the Mel Spectrogram (see Figure 9). The Fast Fourier transform has been used to transfer the audio signal from the time domain to the frequency domain by overlapping windowed chunks of the audio stream. To create the spectrogram, the

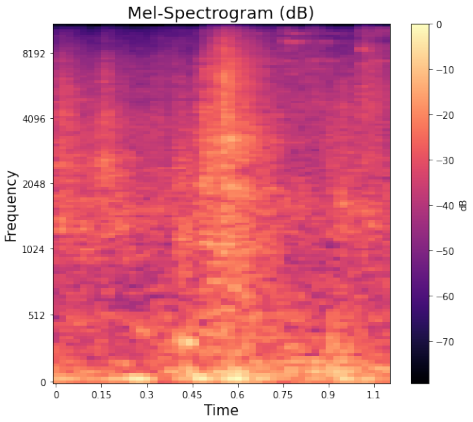

Figure 9: Mel Spectrogram (feature) of an Audio File

y-axis (frequency) is transformed into a log scale and the colour dimension (amplitude) to decibels. Afterward, the Mel Spectrogram is created by mapping the y-axis (frequency) onto the Mel Scale. Then, with ImageGenerator, all pixel values are rescaled from 0-255, so after this step, all the pixel values are in the range of (0,1). Because the spectral power levels of the spectrograms varies greatly even within the same class, this normalizing step is required.

4.5 *Convolutional Neural Network(CNN) Models*

After extracting the feature of the audio files, the Mel spectrograms are fed into the CNN model. This section briefly illustrates two CNN models such as EfficientNet-Bo and ResNet-18 that have been implemented. The process of using pre-trained CNN models for any current task at hand is called transfer learning. Transfer Learning is a process in which models trained on a specific task with a significant quantity of data are applied to a new task to extract valuable features based on its prior knowledge (Iglovikov & Shvets, 2018). In this project, the principle behind transfer learning has been applied, and both the EfficientNet-Bo and ResNet-18 models have been pre-trained on the ImageNet data. After that, they have been executed for the current alarm call classification task. ImageNet is a large visual database that has a total number of 14 million hand-annotated images and it consists of about 20,000 categories or classes (Huh, Agrawal, & Efros, 2016). This large corpus of images is mainly used for object recognition, image classification, and so on. But ImageNet pre-trained CNN models can also be used for audio classification (Palanisamy, Singhania, & Yao, 2020), and the Timm library which is a go-to library to load the pre-trained models on PyTorch is used. Figure 10 shows the entire end-to-end process of the current research methodology.

4.5.1 *EfficientNet-Bo*

EfficientNet-Bo is a convolutional neural network design and scaling approach that uses a compound coefficient to scale all depth/width/resolution dimensions evenly (Tan & Le, 2019).

Three types of layers which are Convolution, Batch-Normalization and Activation have been used in the Efficient-Bo CNN model for the project. Batch normalization continually normalizes the output from the previous layer before passing it on to the next layer and this helps the neural network to stabilize. Conv2D layers make up all these three layers in the hidden layers. Apart from it, a channel attention mechanism, called squeeze-and-excitation (SE) has been applied in the Convolution layers that help re-weighing the information which is present across feature maps with the goal of improving the quality of feature representations. And these convolution layers deal with our two-dimensional matrices as input pictures (Mel Spectrogram). The inputs of the model are 128*128*3 (with the 3 signifying that the images are colored) pixels Mel Spectrogram images and the corresponding category labels. For the convolution, the kernel size equals the size of the filter matrix. Three different kernels or filters of size 1x1, 3x3, and 5x5 have been used in different convolution layers in

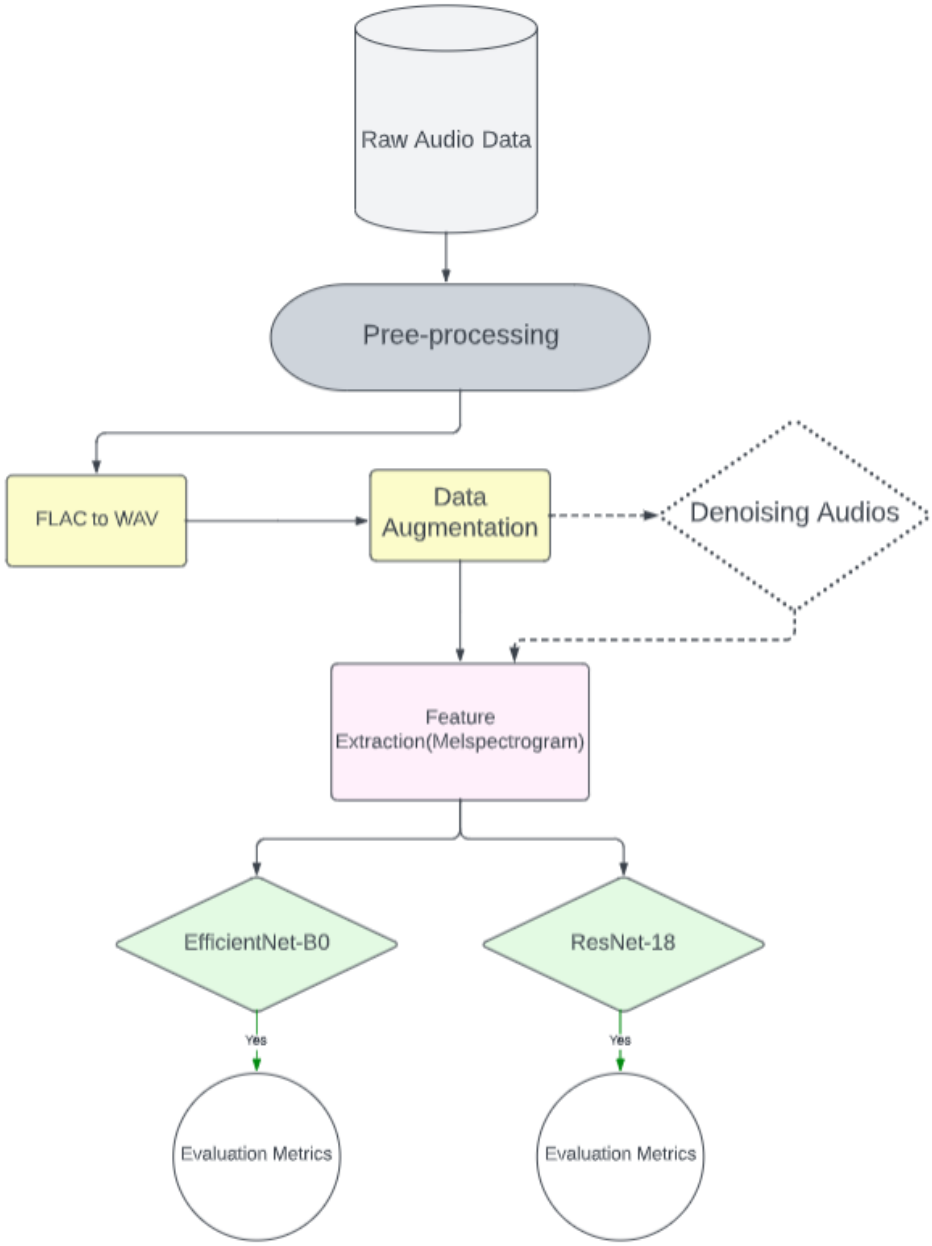

Figure 10: Current Research Pipeline

the Efficient-Bo model. Similarly, padding and stride of size (1,1) and (2,2) have also been applied.

In EfficientNet-Bo, there are 18 blocks and 237 layers if Convolution, Batch Normalization, Pooling, and so on are taken into consideration. Depending on the size of the dataset, this amount might be modified to higher or lower. The loss between the predicted value and the real value of the

model is frequently calculated using cross-entropy loss. CrossEntropyLoss has been used as the loss function. The SiLU, or Sigmoid Linear Units, is the activation function that has been used for the hidden layers, and for the final layer, the output of the torch.nn.Linear layer is fed to the SoftMax function of the cross-entropy loss by using CrossEntropyLoss. Global Pooling has been used to replace Flattening by reducing the dimensionality of feature maps produced by the convolutional layers. For compiling the model, the 'Adam' optimizer has been used and the 'accuracy' metric has been used as the performance metric for the validation set. The learning rate is fixed at 0.001 and the batch size is 32. The momentum has been kept at 0.1. Momentum is basically a variant of Stochastic Gradient Descent and it helps speed up the learning process in the network and helps not get stuck in the local minima.

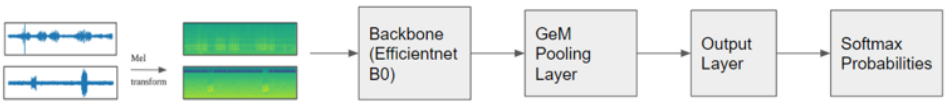

Figure 11: Audio Classification in EfficientNet-B0

Then, the model is trained by splitting the data into training, validation, and test sets with a proportion of 0.7, 0.2, and 0.1 (unseen data by the model) respectively. The number of epochs determines how many times the model cycles over the data. The model no longer improves with each epoch after a certain point. Given the huge number of trainable parameters, the number of epochs is set to 10 for this research work. Afterward, the model gives an accuracy score on the test data.

4.5.2 *ResNet-18*

Residual Network or ResNet is such a Convolutional Neural Network(CNN) that is based on skipping some connections or making a shortcut (Sarwinda, Paradisa, Bustamam, & Anggia, 2021).

In each module of ResNet-18, there are four convolutional layers (excluding the 1x1 convolutional layer). There are a total of 18 layers, including the initial convolutional layer and the last fully-connected layer. As a result, this model is known as ResNet-18. There are 64, 128, 256, and 512 input channels for the first, second, third, and fourth convolutional layers respectively. Apart from it, in each convolutional layer, there are two blocks. Three types of layers which are Convolution, Batch-Normalization and Activation have been used in the ResNet-18 CNN model in this project. Similar to the EfficientNet-B0, these convolution layers deal with our two-dimensional matrices as input pictures (Mel Spectrogram) and take them as

the feature for the model. The inputs of the model are 128*128*3 (with the 3 signifying that the images are colored) pixels Mel Spectrogram images and the corresponding category labels. For the convolution, the kernel size equals the size of the filter matrix. Three different kernels or filters of size 1x1, 3x3, and 7x7 have been used in different convolution layers in the ResNet-18 model. Similarly, padding of size (1,1) and (3, 3) and stride of size (1,1) and (2,2) have also been applied. The loss between the predicted value and the real value of the model is frequently calculated using cross-entropy loss. CrossEntropyLoss has been used as the loss function. The ReLU, or Rectified Linear Units, is the activation function that has been used for the hidden layers and for the final layer, the output of the torch.nn.Linear layer is fed to the SoftMax function of the cross-entropy loss by using CrossEntropyLoss as it is in EfficientNet-B0 as well. The batch size is kept at 32 and Global Pooling has been used to decrease the dimensionality of the feature maps. Just like in the previous model, the Momentum is set to 0.1 which prevents the zigzag movement of the gradient in the weight space. Thus, it serves the purpose of the learning process during the training quite fast. For compiling the model, the 'Adam'

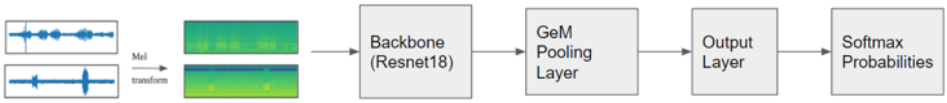

Figure 12: Audio Classification in ResNet-18

optimizer has been used and the 'accuracy' metric has been used as the performance metric for the validation set. The learning rate is fixed at 0.001. Then, the model is trained by splitting the data into training, validation, and test sets with a proportion of 0.7, 0.2, and 0.1 (unseen data in the model) respectively. Similar to EfficientNet-B0, the ResNet-18 model is trained by splitting the data into training, validation, and test sets with a proportion of 0.7, 0.2, and 0.1 (unseen data by the model) respectively. The number of epochs determines how many times the model cycles over the data. The model no longer improves with each epoch after a certain point. The number of epochs is set to 10 for this research work because of a large number of parameters. At last, the model gives an accuracy score on the test data.

4.6 *Evaluation Metrics*

As the dataset is imbalanced, several evaluation metrics have been adopted to measure the performance of the CNN models on the test data. The per-

formance metrics that have been used in this thesis are accuracy, confusion matrix, precision, and F1 score.

5 RESULTS

In this section, the results of the alarm call classification performance of the CNN models such as EfficientNet-B0 and ResNet-18 described in section 4.5 will be illustrated.

As the problem at hand is a classification problem, 'accuracy' has been used as the primary evaluation metric. But the audio dataset is unbalanced i.e. one class has more representation over other classes in the entire dataset, and evaluating the model performances solely on 'accuracy' is not a good indication of evaluation. It is for this reason that 'precision' has also been used as it determines the correctness of our model. In reality, at the farms where there are thousands of pigs, we can afford to miss a few alarm calls (making 'recall' less important for the task) here and there to a small extent, as long as the model is correct (with a high precision score) with a large certainty when it predicts alarm calls. Hence, it is desired for this task to have a high score on the ratio of correct positive predictions to the total positive (precision). But to strike a balance between precision and recall, the 'F1-score' has also been measured and given importance as it combines the precision and recall score into one single metric by taking their harmonic mean.

5.1 *Model performance on noisy data*

As shown in table 2, the overall accuracy, precision, and F1 score for the EfficientNet-B0 performed on the noisy data are 94.29%, 0.9444, and 0.9431 respectively. In other words, a precision score of 0.9444 indicates that when the model predicts a sound as an alarm call or normal call, it is correct 94% of the time. Similarly, the ResNet-18 model gives an overall accuracy of 88.45%, a precision score of 0.81, and an F1 score of 0.883. In figure 13,

Table 2: Alarm Call Classification (with Noisy Data)

CNN Model	Accuracy(%)	Precision	F1-Score
EfficientNet-B0	94.29	0.9444	0.9431
ResNet-18	88.45	0.8102	0.883

the confusion matrix shows that the accuracy of the voice produced by the pig at sneezing call is highest at 96.66%; similarly, it is approximately 92.97% for the normal sound, and 93.21% for the pig's coughing sound.

Particularly, approximately 6.6% and 5.8% of the cases are situations in which the coughing sound and normal pig voices are mistaken for sneezing sounds and coughing, respectively. The most probable reason is that the model has learned some features of coughing sound in such a way that when almost similar features or some of them are present in sneezing, it classifies the coughing sound as sneezing. The same reasoning can also be made when a normal pig sound is classified as coughing. Similary, figure 14 depicts the accuracy of the sound produced by the pig at normal call is highest at 98.01%; similarly, it is approximately 78.05% for the coughing sound, and 89.25% for the pig's sneezing sound. Interestingly, approximately 17.55% of the cases were situations in which the coughing sound is predicted as a sneezing sound.

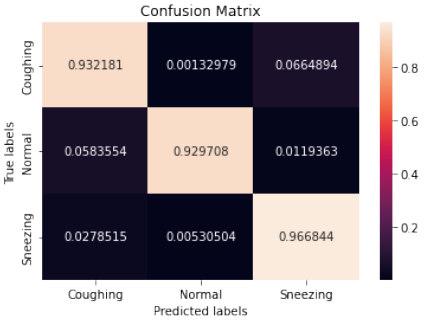

Figure 13: Confusion Matrix for EfficientNet-Bo

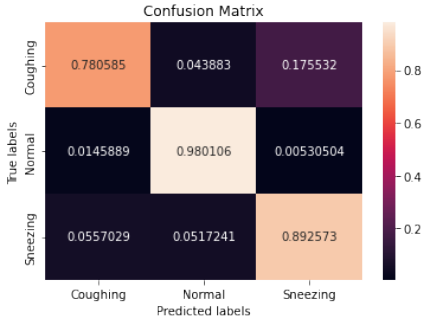

Figure 14: Confusion Matrix for ResNet-18

As both the models have been run over 10 epochs, figure 15 shows the graphical representation of how the loss is changing over epochs for both the train and validation data in the EfficientNet-Bo model. It is evident that the gap between the training loss and validation loss is not big and the loss has been decreasing over epochs and finally stabilises after the epoch number 6. This is indicating a good fit for the model. At the same time, figure 16 depicts that after epoch 4, for the ResNet-18 model, the loss is not changing much and rather stabilises and after epoch 8, the loss on the validation data is increasing again. It is for this reason that the training has been stopped after 10 epochs to avoid overfitting the model.

It is clearly evident from the accuracy scores that the EfficientNet-Bo performs slightly better than the ResNet-18 model in our task as claimed in other related research as well and it is achieved by implementing AutoML property, and reducing parameter size and FLOPs by an order of magnitude (Tan & Le, 2019). Other than that, it can also be concluded from the confusion matrix that the predictive performance of the models differs greatly between classes. Sneezing has the highest accuracy (96.68%) obtained by the EfficientNet-Bo but for the ResNet-18 model,

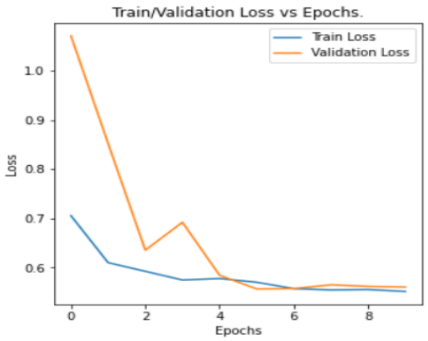

Figure 15: Epoch vs Loss for EfficientNet-B0

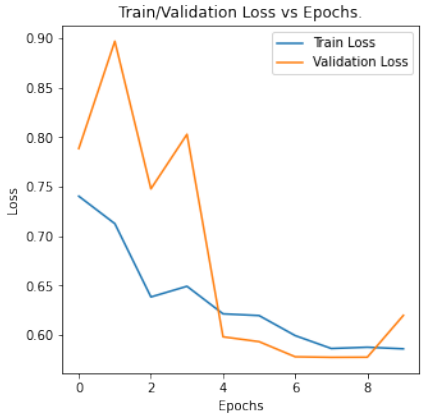

Figure 16: Epoch vs Loss for ResNet-18

normal sound has the top most accuracy of 98.01% among all the three calls. In the ResNet-18, there is a difference of 20% in accuracy when it comes to classifying normal sound and coughing sounds.

5.2 *Model performance on denoised data*

Similarly, to find the answer to the sub-question of the current research, denoised data have also been fed into the model and overall model performances on the denoised data have been recorded. As shown in table 3, the overall accuracy, precision, and F1 score for the EfficientNet-B0 performed on the denoised data are 89.89%, 0.9004, and 0.8879 respectively. Similarly, in the ResNet-18 model, the model obtains an overall accuracy of 77.61%, a precision score of 0.7669, and an F1 score of 0.7403.

Table 3: Alarm Call Classification (with Denoised Data)

CNN Model	Accuracy(%)	Precision	F1-Score
EfficientNet-B0	89.89	0.9004	0.8879
ResNet-18	77.61	0.7669	0.7403

It is obvious from the result (refer to table 3) that both the CNN models have performed poorly with denoised data as compared to how they performed (refer to table 2) with noisy data. It is because the noisy data makes the network less able to recall the training samples, resulting in smaller network weights and a more resilient network with reduced generalization error. In figure 17, the confusion matrix shows that the accuracy achieved by the EfficientNet-B0 at normal call is highest at 97.61%;

similarly, it is approximately 81.38% for the coughing sound and 90.64% for the pig's sneezing sound. Particularly, approximately 15% of the cases in

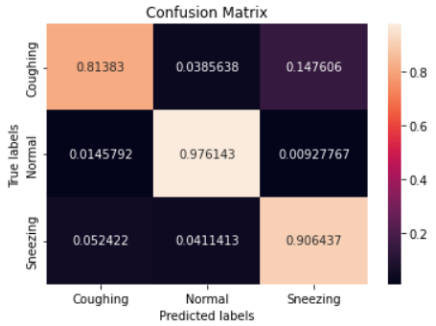

Figure 17: Confusion Matrix for EfficientNet-Bo (Denoised)

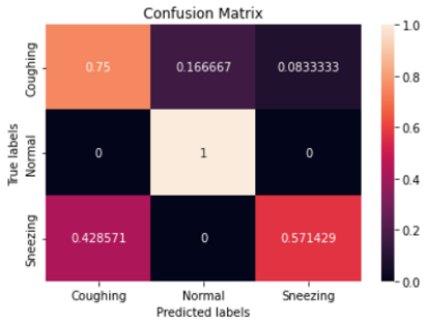

Figure 18: Confusion Matrix for ResNet-18 (Denoised)

which the coughing sound has been wrongly labeled as sneezing. Similary, the figure 18 depicts that the normal call accuracy obtained by the ResNet-18 model is highest at 100%; similarly, it is approximately 75% for the coughing sound and 57.14% for the pig's sneezing sound. The interesting result is that ResNet-18 has achieved a 100% accuracy for the normal call at the cost of obtaining a low accuracy score of 57% for the sneezing call. This occurs as the model is biased towards predicting that normal call class because of the imbalanced dataset; normal sounds are more present in the dataset. This phenomenon brings a pattern of huge class imbalace prediction among the three calls as it is seen from the accuracy percentage in the confusion matrix. Surprisingly, approximately 42% of the cases are such situations in which the sneezing sound is predicted as coughing sounds.

6 DISCUSSION

6.1 *Goal of the research*

The primary goal of this study is to distinguish pig alarm calls such as coughing and sneezing from the normal calls by analyzing the pig sound in the presence of background noise as well as without noise with the help of machine learning models and comparing the model performances against each other. While classifying cattle food anticipation call, estrus call, and so on has already been done in the domain of deep learning for animal welfare (Jung, Kim, Moon, Jhin, et al., 2021). But the task of classifying the sneezing call is such an approach that it has never been attempted for animal welfare. The result of the current study confirmed the societal insights mentioned in the literature which illustrated the importance of the voice of the pig and

the vital information it carries about the animal's extraordinary conditions, such as pain, illness, oestrus, separation from the calf, and hunger or other physical health anomalies (Ikeda & Ishii, 2008; Riede et al., 1997) as well as technically how converting the sound signals into Mel Spectrogram images can be useful for a CNN to obtain a better accuracy in the image recognition field or animal call recognition (Yin et al., 2021). On the basis of the related literature (Palanisamy et al., 2020), this project has implemented two pre-trained Convolutional Neural Networks (CNN) such as Efficient-B0 and ResNet-18. EfficientNet-B0 model achieved an accuracy of 94.29% while ResNet-18 obtained an accuracy score of 88.45% when noise is present in the audio data. This result aligns well with prior research (Xie & Zhu, 2019). And when the noise is removed from the audio files, the accuracy of the models decreases to 89.89% for the EfficientNet-B0 and 77.61% for the ResNet-18 model. This finding contradicts the expectation of the current research and the prior research (Jung, Kim, Moon, Jhin, et al., 2021) where the model performances improve after the background noise is removed. But adding noise to a neural network during training can make it more resilient, resulting in better generalization and faster learning. Keeping the noise to the training process improves its robustness and minimizes generalization error. A low-overhead noise addition approach has shown to be remarkably successful for training very deep architectures. The strategy not only prevents overfitting but can also reduce training loss. Even with a bad initiation, this approach alone permits a fully linked 20-layer deep network to be trained using ordinary gradient descent (Neelakantan et al., 2015). At first glance, noise in data appears to be a formula for making learning more difficult. It's a strange proposal for enhancing performance because noise is supposed to reduce the model's performance during training. But it has been proved in practice that training with noise can increase network generalization (C. M. Bishop et al., 1995). Because training samples change all the time, noisy data makes the network less able to recall them, resulting in smaller network weights and a more resilient network with reduced generalization error. This is the foremost reason why the models implemented in this current research is performing better when there is noise in the data.

Even though between these two pre-trained models, the EfficientNet-B0 CNN model performs slightly better than the ResNet-18 CNN model when it comes to detecting alarm calls, ResNet-18 is still useful as a baseline model and a method of comparison, and above all, confirms that machine learning algorithms are indeed able to classify alarm calls quite accurately.

6.2 *Limitation and Future Work*

Having said these all, there are still limitations of the current study which can be improved for future research. Firstly, in the case of denoised data, the ResNet-18 model obtains an accuracy of 100% for the normal call and performs poorly for classifying the other calls. This is because of an imbalanced dataset where the normal class is more present in number in the dataset and the model is biased towards classifying the normal call more. Class probability estimates attained via supervised learning in imbalanced scenarios systematically underestimate the probabilities for minority class instances, despite good overall performance (Wallace & Dahabreh, 2012). This can be overcome by more data augmentation or adding more data to the minority class. Secondly, even though data augmentation is a well-known method to create more synthetic data from the original data, augmented data might still not hold all the information or all the relevant features of the original data in itself (Ravuri & Vinyals, 2019). The current research worked with a small amount of original data and this is one of the drawbacks of the current research. Thirdly, in the alarm call categorization of pigs, it is possible to discriminate between individual pigs at the farm using deep learning models. Individual pig differences were not taken into account when training the model in this study. Therefore, it is anticipated that future research will produce better findings when these components are analyzed using the video picture. Future work can also involve using the data at hand and training it with pre-trained audio neural networks (PANNs) trained on the large-scale AudioSet dataset which also gives a good performance (Kong et al., 2020). In this way, the performance of models trained on ImageNet like how it is implemented in the current study, and models trained on AudioSet can be compared and selected the one which gives better performance and used for more serious research in the future. Other than it, the EfficientNet-B0 or ResNet-18 models can also be scaled up to B7 for EfficientNet or up to 200 for the ResNet and this is done depth-wise and widthwise. This scaling could gain better results.

That being said, this current research project already captures the alarm calls and accurately classifies them with the help of machine learning algorithms and obtains a meaningful result. And these developed models can be deployed as a web platform in the LIVE setting at farms to give useful information about the physical health of the farm animals to the farm owners so that the farm owners can take prompt actions when anomalies are detected by these deployed models. This will result in saving the lives of the farm animals as well as the maintenance costs of the owners.

7 CONCLUSION

In this study, two deep learning pig alarm call classification models have been developed to determine the physical health status of farm pigs by analysing the vocal sounds of the pigs with machine learning algorithm. In order to develop and implement them, the following research questions have been formulated in the first place as described in the section 2.3:

R1Q *To what extent can a supervised Convolutional Neural Network(CNN) algorithm discriminate between normal calls and alarm calls such as coughing, and sneezing made by the pigs on a rural farm?*

A sub-question has also been constructed, as such:

RQ2 *With how much accuracy can the algorithm in question classify these different calls correctly in the absence of background noise in the audio data?*

Although, the sub-research question is related to the main research question in the sense that in order to find an optimal environment for the algorithm to work the best and implement it in a real-life setting, it is essential to know what that optimal environment is; whether it works better when there is background noise or without noise. While in most of the existing literature, only the approach where background noise was present in the audio data had been implemented, in this research both the approaches which are with noise and without noise have been carried out and compared against each other with a goal to bring novelty to the existing research. The overall accuracy, precision, and F1 score for the EfficientNet-B0 performed on the denoised data are 89.89%, 0.9004, and 0.8879 respectively. Similarly, in the ResNet-18 model, the model obtains an overall accuracy of 77.61%, a precision score of 0.7669, and an F1 score of 0.7403. Lastly, with regard, to the main research question, both the algorithms produced satisfactory results with overall accuracy, precision, and F1 score for the EfficientNet-B0 performed on the noisy data is 94.29%, 0.9444, and 0.9431 respectively. Similarly, the ResNet-18 model gives an overall accuracy of 88.45%, a precision score of 0.81, and an F1 score of 0.883. The voice of animals carries vital information about the animal's extraordinary physical health conditions. Therefore, these research findings bring value to the machine learning approaches to facilitate precision livestock farming (PLF) for farm animal welfare. Also, given the fact that pig is the most consumed meat in the world and global meat consumption is increasing every year, it is quite obvious that more and more farmers will take help of this audio health monitoring system for the farm pigs. These initiatives and results are anticipated to improve animal welfare in the future.

In spite of getting up to the mark results, there is ample room for improvement. Future research can explore using different cepstral features such as Mel frequency cepstral coefficients (MFCCs) and see the difference in performance. Future research could also explore changing the colour maps of Spectrograms and see if it improves the accuracies. Another interesting research subject or sub-topic could be to categorize farm animal 'cry' sounds from normal sounds to monitor their distress by using the same model settings and see if the models can accurately classify the 'cry' sounds.

REFERENCES

Aarnink, A., Swierstra, D., Van den Berg, A., & Speelman, L. (1997). Effect of type of slatted floor and degree of fouling of solid floor on ammonia emission rates from fattening piggeries. *Journal of agricultural engineering research*, *66*(2), 93–102.

Amiriparian, S., Gerczuk, M., Ottl, S., Cummins, N., Freitag, M., Pugachevskiy, S., ... Schuller, B. (2017). Snore sound classification using image-based deep spectrum features.

Bengio, Y., Lamblin, P., Popovici, D., & Larochelle, H. (2006). Greedy layer-wise training of deep networks. *Advances in neural information processing systems*, *19*.

Bishop, C. M., et al. (1995). *Neural networks for pattern recognition*. Oxford university press.

Bishop, J. C., Falzon, G., Trotter, M., Kwan, P., & Meek, P. D. (2019). Livestock vocalisation classification in farm soundscapes. *Computers and Electronics in Agriculture*, *162*, 531–542.

Biswas, T., Pal, C., Mandal, S. B., & Chakrabarti, A. (2014). Audio de-noising by spectral subtraction technique implemented on reconfigurable hardware. In *2014 seventh international conference on contemporary computing (ic3)* (pp. 236–241).

Cannam, C., Landone, C., & Sandler, M. (2010). Sonic visualiser: An open source application for viewing, analysing, and annotating music audio files. In *Proceedings of the 18th acm international conference on multimedia* (pp. 1467–1468).

Chan, W. Y., Cloutier, S., & Newberry, R. C. (2011). Barking pigs: differences in acoustic morphology predict juvenile responses to alarm calls. *Animal Behaviour*, *82*(4), 767–774.

Chedad, A., Moshou, D., Aerts, J.-M., Van Hirtum, A., Ramon, H., & Berckmans, D. (2001). Ap—animal production technology: recognition system for pig cough based on probabilistic neural networks. *Journal of agricultural engineering research*, *79*(4), 449–457.

Chen, H.-M., Huang, C.-J., Chen, Y.-J., Chen, C.-Y., & Chien, S.-Y. (2015). An intelligent nocturnal animal vocalization recognition system. *International Journal of Computer and Communication Engineering*, *4*(1), 39.

Cordeiro, A. F. d. S., Nääs, I. d. A., da Silva Leitão, F., de Almeida, A. C., & de Moura, D. J. (2018). Use of vocalisation to identify sex, age, and distress in pig production. *Biosystems engineering*, *173*, 57–63.

Erhan, D., Manzagol, P.-A., Bengio, Y., Bengio, S., & Vincent, P. (2009). The difficulty of training deep architectures and the effect of unsupervised pre-training. In *Artificial intelligence and statistics* (pp. 153–160).

Gaberson, H. A. (2006). A comprehensive windows tutorial. *Sound and Vibration*, *40*(3), 14–23.

Grandin, T. (2014). Animal welfare and society concerns finding the missing link. *Meat science*, *98*(3), 461–469.

Green, A. C., Clark, C. E., Lomax, S., Favaro, L., & Reby, D. (2020). Context-related variation in the peripartum vocalisations and phonatory behaviours of holstein-friesian dairy cows. *Applied Animal Behaviour Science*, *231*, 105089.

Greiner, L., Stahly, T., & Stabel, T. (2000). Quantitative relationship of systemic virus concentration on growth and immune response in pigs. *Journal of animal science*, *78*(10), 2690–2695.

Hong, M., Ahn, H., Atif, O., Lee, J., Park, D., & Chung, Y. (2020). Field-applicable pig anomaly detection system using vocalization for embedded board implementations. *Applied Sciences*, *10*(19), 6991.

Huang, J., Wang, W., & Zhang, T. (2019). Method for detecting avian influenza disease of chickens based on sound analysis. *Biosystems Engineering*, *180*, 16–24.

Huh, M., Agrawal, P., & Efros, A. A. (2016). What makes imagenet good for transfer learning? *arXiv preprint arXiv:1608.08614*.

Iglovikov, V., & Shvets, A. (2018). Ternausnet: U-net with vgg11 encoder pre-trained on imagenet for image segmentation. *arXiv preprint arXiv:1801.05746*.

Ikeda, Y., & Ishii, Y. (2008). Recognition of two psychological conditions of a single cow by her voice. *Computers and Electronics in Agriculture*, *62*(1), 67–72.

Jung, D.-H., Kim, N. Y., Moon, S. H., Jhin, C., Kim, H.-J., Yang, J.-S., ... Park, S. H. (2021). Deep learning-based cattle vocal classification model and real-time livestock monitoring system with noise filtering. *Animals*, *11*(2), 357.

Jung, D.-H., Kim, N. Y., Moon, S. H., Kim, H. S., Lee, T. S., Yang, J.-S., ... Park, S. H. (2021). Classification of vocalization recordings of laying hens and cattle using convolutional neural network models. *Journal of Biosystems Engineering*, *46*(3), 217–224.

Karam, M., Khazaal, H. F., Aglan, H., & Cole, C. (2014). Noise removal in speech processing using spectral subtraction. *Journal of Signal and Information Processing*, *2014*.

Khamparia, A., Gupta, D., Nguyen, N. G., Khanna, A., Pandey, B., & Tiwari, P. (2019). Sound classification using convolutional neural network and tensor deep stacking network. *IEEE Access*, *7*, 7717–7727.

Kim, J., Caire, G., & Molisch, A. F. (2015). Quality-aware streaming and scheduling for device-to-device video delivery. *IEEE/ACM Transactions on Networking*, *24*(4), 2319–2331.

Kong, Q., Cao, Y., Iqbal, T., Wang, Y., Wang, W., & Plumbley, M. D. (2020). Panns: Large-scale pretrained audio neural networks for audio pattern recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *28*, 2880-2894. doi: 10.1109/TASLP.2020.3030497

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, *25*.

Lee, J., Jin, L., Park, D., & Chung, Y. (2016). Automatic recognition of aggressive behavior in pigs using a kinect depth sensor. *Sensors*, *16*(5), 631.

Lei, Y., Lin, J., He, Z., & Zuo, M. J. (2013). A review on empirical mode decomposition in fault diagnosis of rotating machinery. *Mechanical systems and signal processing*, *35*(1-2), 108–126.

McFee, B., Humphrey, E. J., & Bello, J. P. (2015). A software framework for musical data augmentation. In *Ismir* (Vol. 2015, pp. 248–254).

Mead-Hunter, R., Selvey, L. A., Rumchev, K. B., Netto, K. J., & Mullins, B. J. (2019). Noise exposure on mixed grain and livestock farms in western australia. *Annals of work exposures and health*, *63*(3), 305–315.

Meen, G., Schellekens, M., Slegers, M., Leenders, N., van Erp-van der Kooij, E., & Noldus, L. P. (2015). Sound analysis in dairy cattle vocalisation as a potential welfare monitor. *Computers and Electronics in Agriculture*, *118*, 111–115.

Moynagh, J. (2000). Eu regulation and consumer demand for animal welfare.

Murphy, E., Nordquist, R. E., & van der Staay, F. J. (2014). A review of behavioural methods to study emotion and mood in pigs, sus scrofa. *Applied Animal Behaviour Science*, *159*, 9–28.

Nanni, L., Rigo, A., Lumini, A., & Brahnam, S. (2020). Spectrogram classification using dissimilarity space. *Applied Sciences*, *10*(12), 4176.

Nasirahmadi, A., Edwards, S. A., & Sturm, B. (2017). Implementation of machine vision for detecting behaviour of cattle and pigs. *Livestock Science*, *202*, 25–38.

Neelakantan, A., Vilnis, L., Le, Q. V., Sutskever, I., Kaiser, L., Kurach, K., & Martens, J. (2015). Adding gradient noise improves learning for very deep networks. *arXiv preprint arXiv:1511.06807*.

Oikarinen, T., Srinivasan, K., Meisner, O., Hyman, J. B., Parmar, S., Fanucci-Kiss, A., ... Feng, G. (2019). Erratum: Deep convolutional network for animal sound classification and source attribution using dual audio recordings [j. acoust. soc. am. 145, 654 (2019)]. *The Journal of the Acoustical Society of America*, *145*(4), 2209–2209.

Palanisamy, K., Singhania, D., & Yao, A. (2020). Rethinking cnn models for audio classification. *arXiv preprint arXiv:2007.11154*.

Pijpers, A., Schoevers, E., Van Gogh, H., van Leengoed, L., Visser, I., van Miert, A., & Verheijden, J. (1991). The influence of disease on feed and water consumption and on pharmacokinetics of orally administered oxytetracycline in pigs. *Journal of Animal Science*, *69*(7), 2947–2954.

Price, T., Wadewitz, P., Cheney, D., Seyfarth, R., Hammerschmidt, K., & Fischer, J. (2015). Vervets revisited: A quantitative analysis of alarm call structure and context specificity. *Scientific reports*, *5*(1), 1–11.

Rabiner, L. R. (1978). *Digital processing of speech signals*. Pearson Education India.

Ravuri, S., & Vinyals, O. (2019). Seeing is not necessarily believing: Limitations of biggans for data augmentation.

Riede, T., Tembrock, G., Herzel, H., & Brunnberg, L. (1997). *Vocalization as an indicator for disorders in mammals* (Unpublished doctoral dissertation). Acoustical Society of America.

Sadeghi, M., Banakar, A., Khazaee, M., & Soleimani, M. (2015). An intelligent procedure for the detection and classification of chickens infected by clostridium perfringens based on their vocalization. *Brazilian Journal of Poultry Science*, *17*, 537–544.

Sankupellay, M., & Konovalov, D. (2018). Bird call recognition using deep convolutional neural network, resnet-50. In *Proceedings of acoustics* (Vol. 7, pp. 1–8).

Sarwinda, D., Paradisa, R. H., Bustamam, A., & Anggia, P. (2021). Deep learning in image classification using residual network (resnet) variants for detection of colorectal cancer. *Procedia Computer Science*, *179*, 423–431.

Sauvé, C. C., Beauplet, G., Hammill, M. O., & Charrier, I. (2015). Mother–pup vocal recognition in harbour seals: influence of maternal behaviour, pup voice and habitat sound properties. *Animal behaviour*, *105*, 109–120.

Simard, P. Y., Steinkraus, D., Platt, J. C., et al. (2003). Best practices for convolutional neural networks applied to visual document analysis. In *Icdar* (Vol. 3).

Tan, M., & Le, Q. (2019). Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning* (pp. 6105–6114).

Thornton, B. (2019). Audio recognition using mel spectrograms and convolution neural networks.

Thornton, P. K. (2010). Livestock production: recent trends, future prospects. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *365*(1554), 2853–2867.

Ul Haq, H. F. D., Ismail, R., Ismail, S., Purnama, S. R., Warsito, B., Setiawan, J. D., & Wibowo, A. (2021). Efficientnet optimization on

heartbeats sound classification. In *2021 5th international conference on informatics and computational sciences (icicos)* (p. 216-221). doi: 10.1109/ICICoS53627.2021.9651818

Vandermeulen, J., Bahr, C., Tullo, E., Fontana, I., Ott, S., Kashiha, M., ... others (2015). Discerning pig screams in production environments. *PLoS One*, *10*(4), e0123111.

Van Hirtum, A., Guarino, M., Costa, A., Jans, P., Ghesquiere, K., Aerts, J.-M., ... Berckmans, D. (2003). Automatic detection of chronic pig coughing from continuous registration in field situations. In *Maveba* (pp. 251–254).

Wallace, B. C., & Dahabreh, I. J. (2012). Class probability estimates are unreliable for imbalanced data (and how to fix them). In *2012 ieee 12th international conference on data mining* (pp. 695–704).

Wathes, C., Jones, J., Kristensen, H., Jones, E., & Webster, A. (2002). Aversion of pigs and domestic fowl to atmospheric ammonia. *Transactions of the ASAE*, *45*(5), 1605.

Xie, J., & Zhu, M. (2019). Handcrafted features and late fusion with deep learning for bird sound classification. *Ecological Informatics*, *52*, 74–81.

Xu, W., Zhang, X., Yao, L., Xue, W., & Wei, B. (2020). A multi-view cnn-based acoustic classification system for automatic animal species identification. *Ad Hoc Networks*, *102*, 102115.

Xuan, C., Ma, Y., Wu, P., Zhang, L., Hao, M., & Zhang, X. (2016). Behavior classification and recognition for facility breeding sheep based on acoustic signal weighted feature. *Transactions of the Chinese Society of Agricultural Engineering*, *32*(19), 195–202.

Yeon, S. C., Jeon, J. H., Houpt, K. A., Chang, H. H., Lee, H. C., & Lee, H. J. (2006). Acoustic features of vocalizations of korean native cows (bos taurus coreanea) in two different conditions. *Applied animal behaviour science*, *101*(1-2), 1–9.

Yin, Y., Tu, D., Shen, W., & Bao, J. (2021). Recognition of sick pig cough sounds based on convolutional neural network in field situations. *Information Processing in Agriculture*, *8*(3), 369–379.

Zabir, M., Fazira, N., Ibrahim, Z., & Sabri, N. (2018). Evaluation of pre-trained convolutional neural network models for object recognition. *International Journal of Engineering and Technology*, *7*(3.15), 95–98.