Data in Brief 54 (2024) 110388

Contents lists available at ScienceDirect

Data in Brief

journal homepage: www.elsevier.com/locate/dib

Data Article

SalmonScan: A novel image dataset for machine learning and deep learning analysis in fish disease detection in aquaculture

Md Shoaib Ahmed [a,b,*], Samiha Maisha Jeba [c]

[a] Department of Computer Science, Boise State University, ID, USA
[b] Department of Computer Science and Engineering, Jahangirnagar University, Dhaka, Bangladesh
[c] Department of Computer Science and Engineering, International University of Business Agriculture and Technology, Dhaka, Bangladesh

ARTICLE INFO

Article history:
Received 27 February 2024
Revised 31 March 2024
Accepted 3 April 2024
Available online 6 April 2024

Dataset link: SalmonScan: A Novel Image Dataset for Machine Learning and Deep Learning Analysis in Fish Disease Detection in Aquaculture (Original data)

Keywords:
Fish disease
Salmon fish disease
Disease in aquaculture
Fish disease dataset
Salmon fish disease dataset
Fish disease detection
Computer vision

ABSTRACT

Fish diseases pose a significant threat to food security in aquaculture, as they can lead to considerable reductions in fish production, quality, and profitability. Globally, salmon aquaculture is the quickest-expanding food production system. Detecting and diagnosing fish diseases in their early stages is essential to prevent the spread of diseases and reduce the negative impact on aquaculture's economy and environment. To serve this purpose, we introduce the SalmonScan dataset, a novel and comprehensive collection of images of healthy and infected salmon fish, which can be used for various applications in computer science and aquaculture. Images from online sources and aquaculture salmon firms were gathered to create the dataset. The dataset was then labeled based on the health status of the fish, fresh or infected. Data augmentation methods like rotation, cropping, flipping, and scaling were used to guarantee the dataset's strength and size. The dataset includes 456 images of fresh fish and 752 images of infected fish, both varied and inclusive while maintaining excellent quality. Other researchers and practitioners can use the dataset we have collected for various purposes. They can use it to create and test new or

* Corresponding author.
E-mail addresses: mdshoaibahmed@u.boisestate.edu, shoaibmehrab011@gmail.com (M.S. Ahmed), jebam615@gmail.com (S.M. Jeba).
Social media: @shoaibmehrab (M.S. Ahmed), @JebaMaisha) (S.M. Jeba)

https://doi.org/10.1016/j.dib.2024.110388

existing machine learning (ML) and deep learning (DL) based computer vision models for identifying, categorizing, counting, and analyzing the behavior and biomass of salmon fish. They can also use it to study how different environmental factors affect the health and growth of salmon fish. Furthermore, they can evaluate the accuracy and performance of different image acquisition and processing methods. Additionally, they can explore the feasibility of using generative adversarial networks (GANs) and transfer learning to improve the training speed and stability of DL models designed for fish detection. This SalmonScan dataset paper describes and documents the dataset in detail, making it publicly available and reusable for the research community.

Specifications Table

Subject	Artificial Intelligence Computer Vision and Pattern Recognition Marine Biology Aquatic Science
Specific subject area	Fish diseases detection and health monitoring in aquaculture.
Type of data	Image
Data collection	We collected this dataset from two primary sources: online sources and aquaculture salmon firms. Online sources include websites, blogs, and forums that provide images of healthy and infected salmon fish. Aquaculture salmon firms produce and sell salmon fish for human consumption or other purposes. We contacted several aquaculture salmon firms worldwide and obtained their permission to use their images of salmon fish for our research. We then downloaded and labelled the images according to their health status: fresh or infected. We ensured that our dataset's images have sufficient quality, diversity, and representativeness.
Data source location	Dhaka, Bangladesh
Data accessibility	Repository name: Mendeley DataData identification number (DOI): https://doi.org/10.17632/x3fz2nfm4w.1 Direct URL to data: https://data.mendeley.com/datasets/x3fz2nfm4w/1 Instructions for accessing these data: 1. In the Mendeley data repository, there is a section called 'Files' that contains a direct download link.
Related research article	[1] Ahmed, M. S., Aurpa, T. T., & Azad, M. A. K. (2022). Fish disease detection using image based machine learning technique in aquaculture. *Journal of King Saud University-Computer and Information Sciences, 34*(8), 5170–5182.

1. Value of the Data

- This dataset offers an annotated salmon image collection, facilitating the development of algorithms for automated disease detection.
- Researchers in computer science can use this dataset to evaluate image processing, machine learning, and deep learning models for detecting salmon fish diseases.
- Access to such a dataset encourages cooperation between computer scientists and aquaculture specialists. Computer scientists can use their machine learning and image processing skills to tackle aquaculture problems. In contrast, aquaculture experts can offer domain-specific knowledge and advice for developing models.

- Analysis of the SalmonScan dataset can enhance knowledge of salmon diseases' dynamics, prevalence, distribution, and associated risks. This information is valuable for improving aquaculture management strategies.
- Analyzing patterns within the dataset can provide researchers with valuable insights into the factors affecting farmed salmon populations' health and disease susceptibility. Targeted interventions and preventative measures can be developed using these insights to enhance the fish's overall health and well-being.
- This dataset is essential for training and validating machine and deep learning algorithms in aquaculture research. Researchers can use this dataset to develop reliable models for early detection and management of salmon diseases.

2. Background

The original motivation and context behind compiling the SalmonScan dataset are as follows:

- Aquaculture is facing a significant threat in the form of diseases that affect fish. These diseases can result in significant losses in fish production, quality, and profitability, thereby posing a threat to food security. Globally, salmon aquaculture is the fastest-growing food production system, accounting for 70 percent (2.5 million tons) of the market. However, with the susceptibility of salmon fish to various bacterial, viral, and parasitic infections, such as infectious salmon anemia (ISA), salmonid alphavirus (SAV), and sea lice, the industry faces several challenges.
- Early detection and diagnosis of fish diseases is critical in preventing disease spread and reducing aquaculture's economic and environmental impact. However, current methods for detecting fish diseases are often expensive, time-consuming, invasive, and require expert knowledge and laboratory facilities. Therefore, there is an urgent need to develop alternative methods that are fast, accurate, non-invasive, and accessible for both fish farmers and researchers. Not only image processing but also vision-based machine learning, and deep neural network architectures are effective methods for detecting fish diseases through automatic analysis and classification of fish images based on their visual features, including shape, color, texture, and lesions. However, the development and evaluation of these techniques require a large and diverse dataset of fish images, which needs improvement in the literature.
- The SalmonScan dataset aims to fill this gap by providing a novel and comprehensive collection of images of healthy and infected salmon fish, which can be used for various applications in computer science and aquaculture. This dataset describes and documents the dataset in detail, including its creation, collection, processing, structure, content, and potential applications. This dataset paper also presents some preliminary results of using machine learning techniques along with some image processing to the dataset, demonstrating its usefulness and validity. This dataset paper adds value to the original research article by making the dataset publicly available and reusable for other researchers and practitioners in the field.

3. Data Description

In our data repository, we demonstrate both raw and augmented datasets. The SalmonScan dataset (raw) consists of 24 fresh fish and 91 infected fish [Some raw datasets have been accidentally removed due to server cleaning in the past.] On the other hand, the SalmonScan dataset (augmented) comprises around 1208 salmon photos, separated into two categories.

- Fresh Salmon (healthy fish without visible disease signs), with 456 images.
- Infected Salmon is showcasing signs of disease with 752 images.

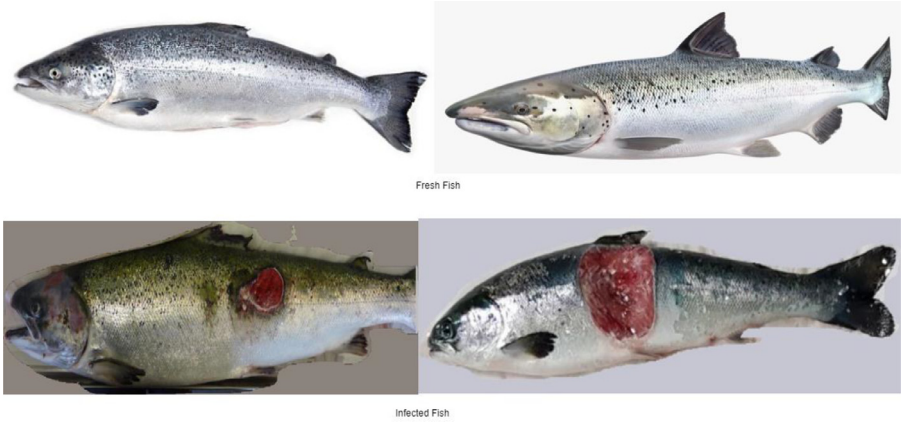

Fig. 1. Two samples of fresh fish and infected fish from SalmonScan dataset.

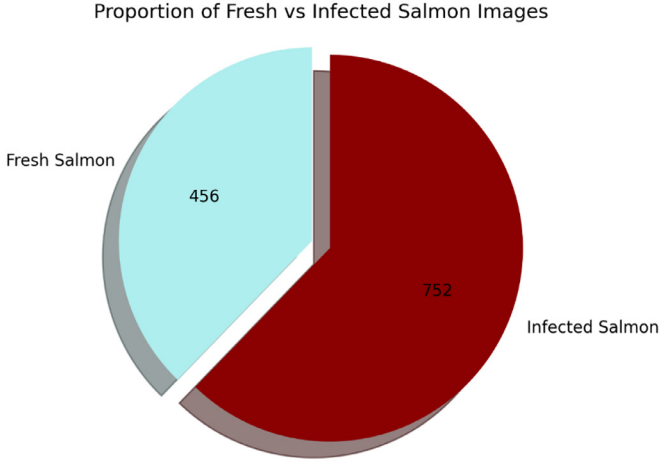

Fig. 2. Percentage of Fish Distribution in the dataset.

Each class offers a representative and diversified collection of images depicting a variety of angles, scales, and lighting situations. The photos have been meticulously curated for various computer vision tasks to guarantee their excellent quality. The sample of both two classes of images is depicted in Fig. 1.

The dataset has been preserved in PNG format and is currently not divided into training, validation, and test sets. The pie chart in Fig. 2 gives an easier-to-view illustration of the ratio of fresh to infected salmon photos. To provide a clear and straightforward comparison, each section of the pie chart reflects the quantity of images for each category.

Nevertheless, a common split would be 20% for validation, 20% for testing, and 60% for training, providing the following counts in Table 1.

A three-dimensional bar chart in Fig. 3 shows each dataset division's total number of fish. The height of each bar indicates the overall count for each split, making it easy to compare the training, validation, and test sets.

Table 1
Common splitting ratio of SalmonScan dataset.

Split	Fresh Fish	Infected Fish	Total
Training	273	451	724
Validation	92	151	243
Test	91	150	241
Total	456	752	1208

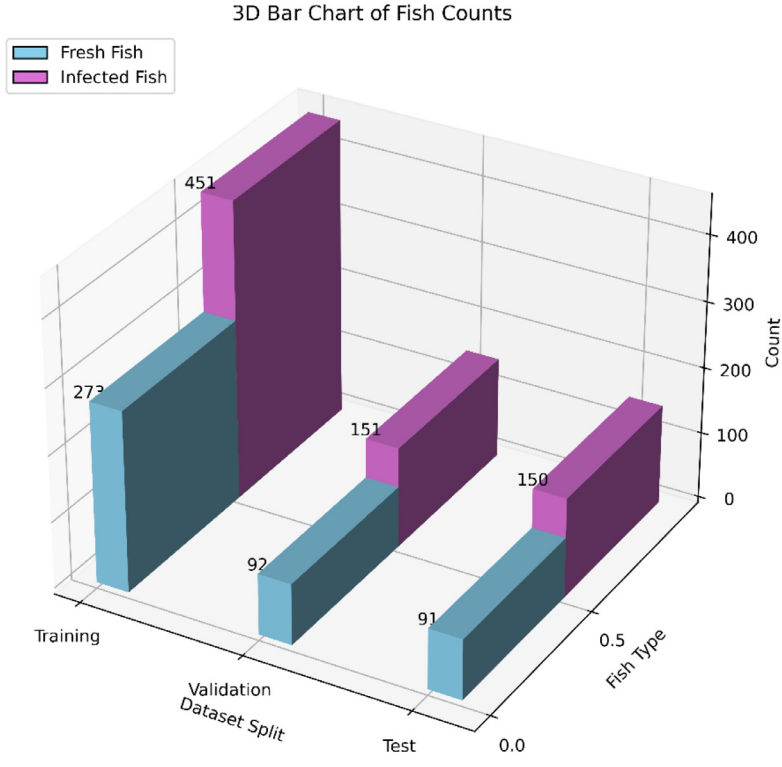

Fig. 3. Total number of fish in each training, testing, and validation division.

4. Experimental Design, Materials and Methods

Experimental Setup:

We used a local computer with 16 GB of RAM and an Intel(R) Core (TM) i7-1100 CPU is utilized to collect data and preprocessing. Machine learning models are trained via the cloud-based notebook service Google Colab. With Ubuntu operating system compatibility, the machine has GPU and TPU capabilities. It is especially compatible with the NVIDIA-manufactured Tesla K-80 GPU, which has two gigabytes of GPU memory.

Data Preprocessing:

The input images were pre-processed to improve their quality and appropriateness for future investigation. The subsequent actions were performed below, and Fig. 4 illustrates the process.

1. **Resizing:** Resizing images to uniform dimensions is an essential preprocessing step in image-based machine-learning tasks. It guarantees that the width and height of each input image are the same, which makes it easier to analyze data consistently through a

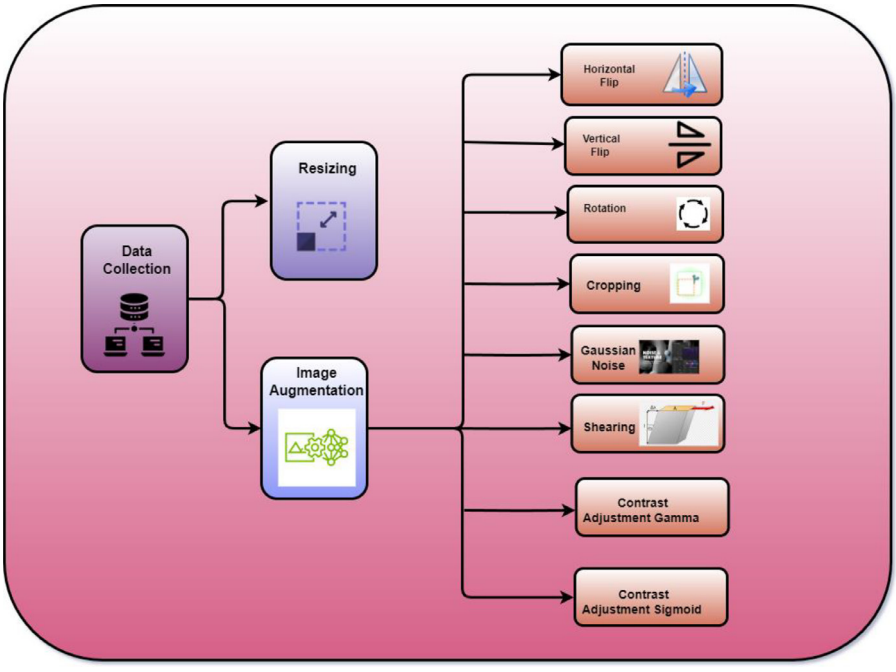

Fig. 4. The procedure of data preprocessing.

model. Here, in order to guarantee compliance with the learning process, every image was shrunk to a consistent 600 pixels wide by 250 pixels of height.

2. **Image Augmentation:** Image augmentation is a technique for artificially expanding the dataset. This is useful when presented with a data collection that has relatively few data samples. This is problematic for deep learning since the model tends to overfit when trained on a small number of data samples. Several image augmentation techniques were used for the input images to overcome the small number of images. Among them were:
    a. **Horizontal Flip:** Horizontal flip is a data augmentation technique that flips the rows and columns of a matrix horizontally. This can help enhance a model's accuracy by exposing it to different variants of the same images.
    b. **Vertical Flip:** Vertical flip is a data augmentation method that involves vertically flipping a matrix's rows and columns. An image upside down along the x-axis will appear consequently.
    c. **Rotation:** Rotation is one popular method of data augmentation. The item in the original picture is randomly rotated by a certain number of degrees, either clockwise or anticlockwise, to alter its location inside the frame.
    d. **Cropping:** Random crop is a powerful data augmentation technique utilized in image processing, particularly beneficial when training deep learning models to recognize fish in various environments. We enhance the model's generalization ability by generating random subsets of the original fish images, thereby improving its performance.
    e. **Gaussian Noise:** Gaussian noise is frequently employed in image augmentation to generate random fluctuations in pixel intensity levels, simulating the inherent variability found in real-world photographs. It improves the model's capacity to generalize new data and helps avoid overfitting. Each pixel receives an individual addition of Gaussian noise drawn from a normal distribution.

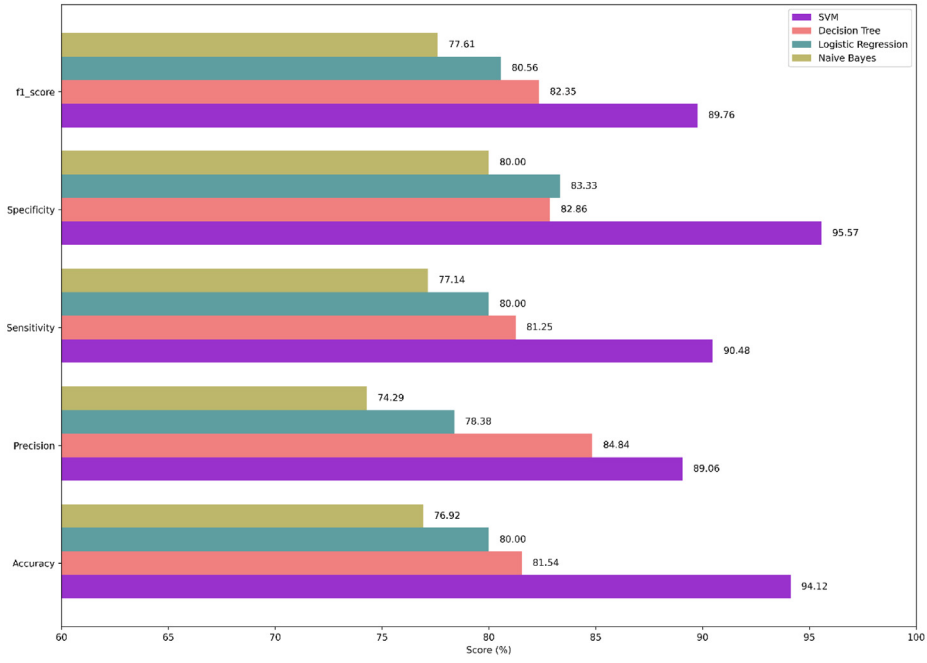

Fig. 5. Comparison of classifiers evaluation metric.

The equation to add Gaussian noise to an image is:

$$I_{\text{noisy}}(x, y) = I(x, y) + N\left(0, \sigma^2\right)$$

- $I_{\text{noisy}}(x, y)$ is the pixel intensity at coordinates (x, y) in the noisy image.
- $I(x, y)$ is the original pixel intensity at coordinates (x, y) in the input image.
- $N(0, \sigma^2)$ is a random sample drawn from a normal (Gaussian) distribution with mean 0 and variance σ^2, representing the amount of noise to be added.

f. **Shearing:** Shear transformation is the process of slanting the picture. Shearing differs from rotation in that instead of rotating the picture, we set one axis and stretch it to a certain angle called the shearing angle. This is a type of stretching that isn't noticeable when rotating. A counter-clockwise shear angle expressed in degrees will be represented by the value of the shear range, which will be float.

g. **Contrast Adjustment (Gamma):** A popular method in image processing and augmentation is contrast modification, which is frequently accomplished by gamma correction. By adjusting the pixel values according to a non-linear mapping function, gamma correction modifies the brightness and contrast of a picture.
To modify the contrast of the picture, scale the pixel values to $255 * ((\text{v}/255) ** \text{gamma})$. Gamma values between 0.5 and 2.0 appear to be reasonable and v is the pixel value.

h. **Contrast Adjustment (Sigmoid):** Another image processing method that seeks to improve or alter contrast is contrast adjustment via the use of a sigmoid function. This approach modifies the contrast by applying a sigmoid function to the pixel intensities.
Adjusted Pixel Intensity $= 255 * 1/(1 + \exp(\text{gain} * (\text{cutoff} - v/255)))$
  - v denotes the pixel intensity value (which ranges from 0 to 255).
  - Gain is the parameter that determines the steepness of the sigmoid curve. It is consistently and randomly selected from the interval [3, 10].

- The cutoff parameter sets the middle of a sigmoid curve. Additionally, a uniform random sample is taken from the interval [0.4, 0.6].

4.1. Data evaluation

We used the SalmonScan dataset to evaluate the performance of our proposed image processing and machine learning strategies [1]. We utilized various machine-learning algorithms, for instance, SVM, Decision Tree, Logistic Regression, and Naive Bayes; from these, the SVM [2] outperformed others. Fig. 5 illustrates the result of our SalmonScan dataset, where we consider some evaluation metrics such as accuracy, precision, f1-score, sensitivity, and specificity of our utilized algorithms.

Limitations

The SalmonScan dataset has some limitations that should be acknowledged and addressed in future work. One of the main limitations is the low amount of image data, especially for the fresh fish class. The dataset contains only 456 images of fresh fish and 752 images of infected fish, which may need more to capture the total variability and diversity of salmon fish in different conditions and environments. The low amount of image data also poses challenges for training and testing deep learning models, which usually require large and balanced datasets to achieve good performance and generalization. The low amount of image data is because it took much work to find and collect images of salmon fish online and offline. Online sources often have low-quality, noisy, or irrelevant images, while offline sources require permission, cooperation, and access to aquaculture salmon firms. Therefore, we suggest that future work should focus on increasing the size and quality of the SalmonScan dataset by using more online and offline sources.

Ethics Statement

The authors of this paper are aware of the ethical statements of this journal, and they agree with it.

Data Availability

SalmonScan: A Novel Image Dataset for Machine Learning and Deep Learning Analysis in Fish Disease Detection in Aquaculture (Original data) (Mendeley Data).

CRediT Author Statement

Md Shoaib Ahmed: Conceptualization, Formal analysis, Writing – original draft, Writing – review & editing, Data curation, Visualization, Supervision; **Samiha Maisha Jeba:** Writing – original draft, Visualization, Writing – review & editing.

Acknowledgements

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

I, Md Shoaib Ahmed, would like to express my deepest gratitude to Tanjim Taharat Aurpa for her unwavering support and love throughout my research journey. Her encouragement and patience have been invaluable to me. Tanjim Taharat Aurpa, will you marry me?

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] M.S. Ahmed, T.T. Aurpa, M.A.K. Azad, Fish disease detection using image based machine learning technique in aquaculture, J. King Saud Univ. 34 (8) (2022) 5170–5182.
[2] S. Suthaharan, S. Suthaharan, Support vector machine, in: Machine Learning Models and Algorithms for Big Data Classification: Thinking with Examples for Effective Learning, 2016, pp. 207–235.