(RESEARCH ARTICLE)

Grasserie disease detection in silkworm using machine learning methods

Manju G *

Department of Computer Science, Panampilly Memorial Government College, Chalakudy, Kerala, India.

Abstract

Grasserie disease, caused by the Bombyx mori nuclear polyhedrosis virus (BmNPV), is one of the most severe diseases affecting silkworms, leading to significant economic losses in sericulture. This study leverages machine learning techniques to detect Grasserie disease in FC1 x FC2 double hybrid silkworms, using image processing and classification methods to automate detection. A dataset of 737 images, consisting of 322 infected and 415 healthy silkworms, was analyzed. Local Binary Patterns (LBP) were used for feature extraction, followed by Principal Component Analysis (PCA) for dimensionality reduction. A decision tree classifier achieved an accuracy of 92.1%, recall of 92.4%, precision of 90.2%, F1 score of 91.6%, and specificity of 90.7%. This paper outlines the methodology, discusses the results, and suggests future improvements for practical applications in sericulture.

Keywords: Silkworm; Grasserie; Bombyx mori; Machine learning; Decision Tree

1. Introduction

Grasserie disease, caused by the Bombyx mori nuclear polyhedrosis virus (BmNPV), represents a critical challenge for the silk industry. Affecting silkworm larvae, this disease often leads to significant decreases in silk quality and yield, impacting the economy in regions heavily dependent on sericulture [1]. Grasserie-infected silkworms exhibit symptoms such as swollen bodies and exude a milky white fluid due to virus-induced degradation of hemocytes. Traditional detection methods for Grasserie are largely based on visual inspection, requiring trained personnel to identify symptoms manually, which can be both labor-intensive and susceptible to errors, particularly under large-scale sericulture operations. These limitations highlight the need for automated and precise diagnostic systems capable of early detection to prevent disease spread [2-5].

Machine learning, particularly in image processing, has seen transformative applications across agriculture, veterinary sciences, a nd disease diagnosis. For agricultural diseases, machine learning methods have shown promise in detecting infections in plants and animals by analyzing subtle visual and textural changes [6,7]. In the case of Grasserie, these methods can leverage image-based feature extraction techniques to identify specific patterns associated with the disease, enabling quick and objective diagnoses. Through automation, machine learning can also help reduce labour costs and improve diagnostic consistency. While machine learning models have been widely explored for human health diagnostics, their application in sericulture remains relatively underexplored [8-10].

This study focuses on the detection of Grasserie disease in FC1 x FC2 double hybrid silkworms by employing machine learning techniques. Using a dataset comprising 737 silkworm images, we applied Local Binary Patterns (LBP) for feature extraction and Principal Component Analysis (PCA) for dimensionality reduction, followed by a decision tree classifier to distinguish between infected and healthy silkworms. LBP is particularly effective for capturing local textures, which are essential in differentiating Grasserie symptoms [2,4]. PCA further helps reduce dimensionality, simplifying the dataset for efficient classification. The choice of a decision tree classifier aligns with the objective of

* Corresponding author: Manju G

creating an interpretable model, as it provides a clear rationale for each classification decision based on observed features.

The following sections describe the dataset, preprocessing, feature extraction, dimensionality reduction, and classification methods used. We then present the results of our experiments, demonstrating the model's effectiveness and discussing its implications for disease detection in sericulture. Lastly, we conclude with insights into the potential improvements and future applications of machine learning for diagnosing Grasserie and other diseases in silkworms.

2. Methodology

2.1. Dataset

The dataset utilized in this study consists of 737 images of FC1 x FC2 double hybrid silkworms, including 322 infected and 415 healthy specimens. Each image was labeled according to the disease status, allowing the model to differentiate between Grasserie-positive and negative cases. Figure 1 shows the sample images of healthy and Grasserie-infected silkworms, showing visible symptoms like swelling and body colour changes.
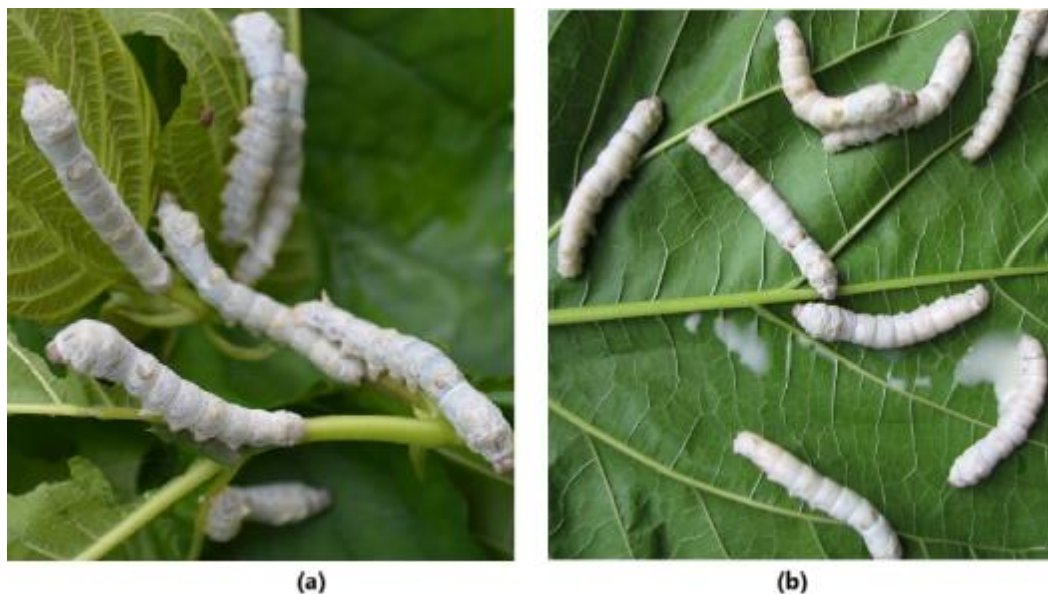

Figure 1 Sample images of (a) healthy and (b) Grasserie-infected silkworms

2.2. Preprocessing

Preprocessing steps were crucial in preparing the images for feature extraction and analysis:

- Grayscale Conversion: To reduce computational load, images were converted to grayscale, simplifying them while retaining essential texture details.
- Normalization: All images were normalized to ensure uniform pixel intensity distribution, facilitating better feature extraction.
- Noise Reduction: A median filter was applied to remove minor artifacts, maintaining the integrity of essential features like body texture and patterns without introducing distortions.

2.2.1. Feature Extraction Using Local Binary Patterns (LBP)

LBP was selected for feature extraction in this study due to its effectiveness in capturing fine-grained textural details, which are essential for distinguishing between healthy and Grasserie-infected silkworms. The LBP method is a powerful tool in texture analysis, as it operates by comparing each pixel's intensity with its surrounding neighbors in a defined pattern, thereby capturing local spatial information [11,12]. This approach makes LBP particularly suited for tasks involving subtle variations in texture, which are often indicative of disease in biological images. The LBP process works by examining each pixel and its surrounding neighborhood, comparing the center pixel's intensity with each neighboring pixel. If a neghboring pixel has intensity greater than or equal to the center pixel, a binary "1" is assigned;

otherwise, a "0" is assigned. This binary sequence, typically 8 bits long when using an 8-neighbor pattern, is then converted into a decimal value. The resulting values across an image form a histogram representing the frequency of different binary patterns, which can be highly descriptive of the underlying texture.

In our study, each silkworm image was converted into an LBP histogram, which served as a compact representation of its textural characteristics. This transformation captured the minute differences in texture associated with the Grasserie infection, such as body swelling and changes in skin texture, which are often too subtle for the human eye to detect consistently. LBP's ability to encode these variations into a binary format allows for efficient analysis, as each image's texture is distilled into a feature vector that preserves critical distinctions indicative of disease. Moreover, LBP has the advantage of being invariant to monotonic changes in grayscale, meaning that variations in lighting or minor changes in image intensity do not affect its performance. This property makes LBP particularly robust for real-world applications, where imaging conditions might vary. Given these attributes, LBP not only provides a reliable way to capture relevant features but also significantly reduces the dimensionality of the data by converting complex textures into simpler numerical forms.

2.2.2. Dimensionality Reduction with Principal Component Analysis (PCA)

PCA was employed to reduce the complexity of the feature set derived from the Local Binary Patterns (LBP) while preserving critical information necessary for effective classification [13]. As a statistical technique, PCA is highly effective for transforming high-dimensional data into a smaller number of dimensions by identifying the principal components, or directions, in which the data varies the most. By focusing on these principal components, PCA allows us to retain the most significant patterns in the data, which are essential for distinguishing between infected and healthy silkworms. The LBP-derived feature set, although highly informative, is inherently high-dimensional due to the wide range of textures it captures from each silkworm image. Such a feature set can be challenging for machine learning algorithms, as high-dimensional data may include redundant or irrelevant features that add noise, reduce model interpretability, and increase computational cost. PCA addresses this by transforming the original features into a set of orthogonal, linearly uncorrelated components, each explaining a portion of the total variance in the data. By ranking these components according to their variance, PCA enables the selection of a subset that encapsulates the majority of the dataset's informational content [14,15].

In our study, PCA was applied to condense the LBP features, effectively reducing the dimensionality of the feature space while preserving the variance that is most relevant to disease detection. This process involved calculating the covariance matrix of the LBP-derived features, followed by identifying the eigenvectors and eigenvalues. The eigenvectors with the largest eigenvalues correspond to the directions in which the data varies the most, and these vectors form the principal components. By retaining only the top principal components, we retained the structure and information of the original data but with significantly fewer dimensions.

2.2.3. Classification with Decision Tree

For the classification of Grasserie disease in silkworms, a decision tree classifier was selected due to its interpretability, simplicity, and effectiveness, particularly in binary classification tasks like this one. Decision trees are non-parametric models that operate by recursively partitioning the feature space into regions with homogeneous labels [16]. At each step, the algorithm selects the feature and threshold that best split the data based on a chosen criterion, such as Gini impurity or entropy. This process results in a hierarchical structure that defines a sequence of decision rules, guiding the model's classification of new samples by traversing these rules from the root node to a leaf node.

In our study, the decision tree model was trained on 80% of the dataset, with the remaining 20% reserved for testing. This split allowed us to evaluate the model's generalization to unseen data, ensuring its reliability and effectiveness when applied to real-world samples. The training process involved recursively selecting LBP-derived, PCA-reduced features that best differentiated between infected and healthy silkworms, creating a tree structure where each internal node represents a decision based on one of these features. The leaf nodes of the tree indicate the predicted class, either "Grasserie-infected" or "healthy," based on the path traversed by a sample through the tree.

To achieve optimal performance, hyperparameters such as the maximum depth of the tree, minimum samples per leaf, and the splitting criterion were tuned through a systematic grid search [17, 18]. The maximum depth parameter controls the complexity of the model by limiting the number of hierarchical splits, which helps prevent overfitting by ensuring the tree does not grow excessively complex. Limiting the tree's depth forces it to focus on the most impactful splits, thereby improving its robustness. The minimum samples per leaf parameter, on the other hand, controls the minimum number of samples required to create a new node. By adjusting this parameter, we ensured that each leaf

node had sufficient data to provide a stable prediction, reducing the likelihood of spurious splits that might capture noise rather than meaningful patterns.

The grid search process systematically evaluated different combinations of these hyperparameters to identify the configuration that maximized classification accuracy and minimized the risk of overfitting [19]. For this purpose, a range of values for each parameter was defined, and the model's performance was assessed on a validation set for each possible combination. This method allowed us to select the best-performing hyperparameters based on metrics such as accuracy, precision, recall, and F1 score, ensuring that the decision tree could make accurate predictions while retaining interpretability.

3. Results and Discussion

The decision tree classifier achieved the following performance metrics as shown in table 1.

Table 1 Classification results

Model	LBP-PCA-DT
Accuracy	92.1
Recall	92.4
Specificity	90.7
Precision	90.2
f1 score	91.6

The model's performance metrics demonstrate a strong ability to detect Grasserie disease in silkworms accurately and reliably. With an overall accuracy of 92.1%, the model shows that it can classify silkworms as healthy or infected with high precision, a crucial feature for early intervention and disease management in sericulture. The breakdown of individual performance metrics further underscores the robustness and reliability of this detection approach. The recall of 92.4% indicates that the model is highly effective at identifying infected silkworms, minimizing the number of false negatives. This is particularly important in disease detection, where missing an infected sample can lead to uncontrolled disease spread. In this study, a high recall means the model is capturing most cases of infection, supporting timely intervention efforts.

The specificity of 90.7% reflects the model's capacity to correctly identify healthy silkworms, effectively reducing false positives. In practical applications, this high specificity ensures that healthy silkworms are not incorrectly classified as infected, thereby avoiding unnecessary interventions and associated costs. By balancing high specificity with a high recall rate, the model provides a reliable tool for decision-making in the early stages of disease control. Precision, calculated at 90.2%, further supports the model's reliability in classification tasks by indicating that the majority of samples predicted as infected were truly infected. High precision is essential to prevent misclassification, which can otherwise lead to undue treatments or incorrect preventive measures. Paired with a high recall, the precision score assures that the model is both sensitive and specific in detecting Grasserie disease. The F1 score of 91.6% confirms the model's balanced performance across precision and recall, highlighting its overall efficiency in disease detection. The F1 score, a harmonic mean of precision and recall, is a useful indicator in this binary classification task because it shows the model's ability to handle both types of classification errors (false positives and false negatives). This balance is key to ensuring that the model is robust enough for operational use.

Receiver Operating Characteristic (ROC) Curve and Area Under Curve (AUC): An ROC curve was drawn to assess the model's classification threshold performance visually ans shown in figure 2. The curve plots the true positive rate against the false positive rate at various threshold settings, providing insight into the model's ability to separate the infected from the healthy class across different decision boundaries. The AUC score of 0.91 is a strong indicator of the model's discriminative power. An AUC near 1.0 signals that the model effectively differentiates between infected and healthy silkworms, making it a dependable classifier for this task [20]. In field settings, a high AUC enhances confidence in the model's predictions, regardless of where the threshold is set.

The choice of Local Binary Patterns (LBP) for feature extraction played a critical role in the model's success. LBP is known for its robustness in capturing local textural patterns, which are often indicative of Grasserie symptoms. By converting each image into a histogram of LBP features, the model could encode essential textural information compactly, preserving critical differences between infected and healthy silkworm samples. The application of PCA further refined this feature set by reducing dimensionality, which ensured that the decision tree classifier could operate efficiently. PCA selected the most variance-explaining features, removing noise and irrelevant data, which might otherwise have burdened the classifier. This reduced dataset proved sufficient for high performance, allowing the decision tree to focus on informative features alone. The decision tree classifier, chosen for its interpretability and effectiveness in binary classification tasks, yielded impressive results. However, exploring other classifiers such as Support Vector Machines (SVM) or ensemble methods (e.g., Random Forests or Gradient Boosting) could provide insights into further optimization. SVM, for example, may offer increased accuracy in cases with clearly defined margins between classes, while ensemble methods can improve robustness by combining predictions from multiple models.
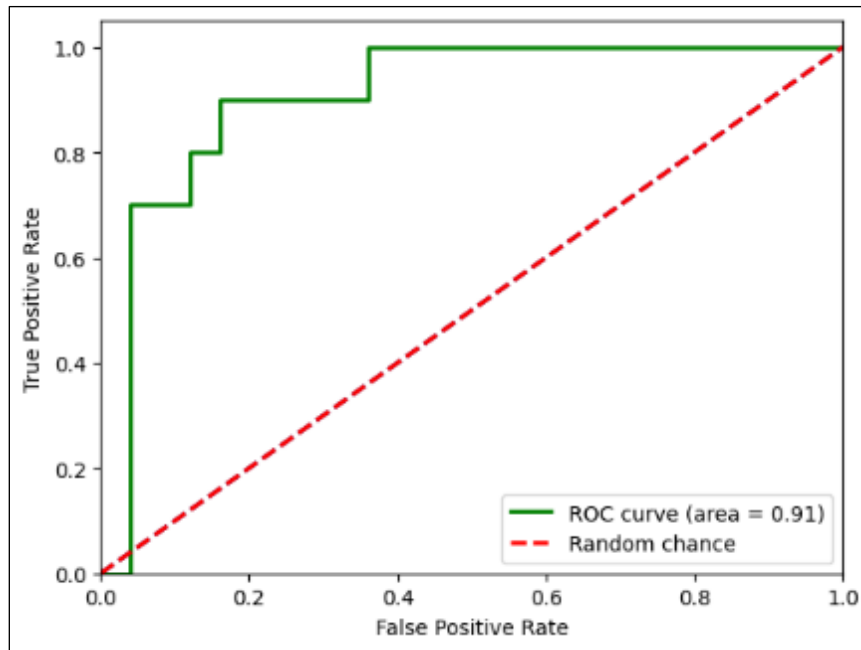

Figure 2 ROC curve of classification model

4. Conclusion

This study demonstrates the feasibility of a machine learning approach to detect Grasserie disease in silkworms, achieving over 90% accuracy with a decision tree classifier. By automating disease detection, our approach provides a reliable alternative to manual inspection, potentially increasing diagnostic speed and accuracy in sericulture. The use of LBP and PCA allowed us to reduce data complexity while retaining essential features for disease classification. Future research could explore expanding the dataset to include other silkworm varieties, environmental conditions, and additional image sources to ensure model generalizability. Ensemble models and deep learning approaches may also be explored to further enhance classification performance.

Incorporating this model into a mobile application could enable field-based, real-time diagnostics, empowering sericulture practitioners with immediate feedback. This approach holds promise for wider applications in agriculture and veterinary sciences, where early disease detection can have substantial economic and ecological benefits.

References

[1]   Sharma A, Sharma P, Thakur J, Murali S, Bali K. Viral diseases of mulberry silkworm, Bombyx mori L.-A review. Journal of Pharmacognosy and Phytochemistry. 2020;9(2S):415-23.

[2]   Thomas S, Thomas J. Nondestructive and cost-effective silkworm, Bombyx mori (Lepidoptera: Bombycidae) cocoon sex classification using machine learning. International Journal of Tropical Insect Science. 2024 Mar 25:1-3.

[3] Nagashetti SM, Biradar S, Dambal SD, Raghavendra CG, Parameshachari BD. Detection of disease in Bombyx mori silkworm by using image analysis approach. In2021 IEEE Mysore Sub Section International Conference (MysuruCon) 2021 Oct 24 (pp. 440-444). IEEE.

[4] Thomas S, Thomas J. Non-destructive silkworm pupa gender classification with X-ray images using ensemble learning. Artificial Intelligence in Agriculture. 2022 Jan 1;6:100-10.

[5] Singla S, Garg S, Garg I, Jha TK, Singh B, Arya H. Disease Detection in Bombyx Mori Silkworm Using Deep Learning Algorithm CNN. In2023 International Conference on Advanced Computing & Communication Technologies (ICACCTech) 2023 Dec 23 (pp. 316-320). IEEE.

[6] Shilpashree PS, Karegowda AG, Suresh KV, Rani L. Classification of Healthy and Diseased Silkworms using Ensemble Learning and CNN. In2023 International Conference on Smart Systems for applications in Electrical Sciences (ICSSES) 2023 Jul 7 (pp. 1-5). IEEE.

[7] Thomas S, Thomas J. A review on existing methods and classification algorithms used for sex determination of silkworm in sericulture. InInternational Conference on Intelligent Systems Design and Applications 2020 Dec 12 (pp. 567-579). Cham: Springer International Publishing.

[8] Binson VA, George MM, Sibichan MA, Raj M, Prasad K. Freshness evaluation of beef using MOS based E-Nose. In2023 International Conference on Intelligent Data Communication Technologies and Internet of Things (IDCIoT) 2023 Jan 5 (pp. 792-797). IEEE.

[9] Thomas S, Thomas J. Comparative Analysis of Non-Destructive Silkworm Cocoon Sex Classification using Machine Learning Models Based on X-Ray and Camera Images. In2024 International Conference on Intelligent Algorithms for Computational Intelligence Systems (IACIS) 2024 Aug 23 (pp. 1-7). IEEE.

[10] Gárate-Escamila AK, El Hassani AH, Andrès E. Classification models for heart disease prediction using feature selection and PCA. Informatics in Medicine Unlocked. 2020 Jan 1;19:100330.

[11] Thomas S, Thomas J. An optimized method for mulberry silkworm, Bombyx mori (Bombycidae: Lepidoptera) sex classification using TLBPSGA-RFEXGBoost. Biology Open. 2024 Jun 14:bio-060468.

[12] Binson VA, Thomas S. The Development of a Mobile E-Nose System for Real-Time Beef Quality Monitoring and Spoilage Detection. Engineering Proceedings. 2023 Nov 9;56(1):256.

[13] Kona VS, Subramoniam M, Binson VA. An Efficient Smart-IoT-Aided Waste Management System with Optimal Shortest Path Routing Using Mud Ring Algorithm. In2023 IEEE International Conference on Recent Advances in Systems Science and Engineering (RASSE) 2023 Nov 8 (pp. 1-7). IEEE.

[14] Nagavelli U, Samanta D, Chakraborty P. Machine Learning Technology-Based Heart Disease Detection Models. Journal of Healthcare Engineering. 2022;2022(1):7351061.

[15] Binson VA, Thomas S, Manju G, KS SK. Precision crop prediction using IoT-enabled soil sensors and Machine Learning. ITEGAM-JETIA. 2024 Sep 25;10(49):65-70.

[16] Ouhami M, Hafiane A, Es-Saady Y, El Hajji M, Canals R. Computer vision, IoT and data fusion for crop disease detection using machine learning: A survey and ongoing research. Remote Sensing. 2021 Jun 25;13(13):2486.

[17] KS SK, VA B. An IoT-Enabled Real-Time Crop Prediction System Using Soil Fertility Analysis. Eng. 2024 Oct 8;5(4):2496-510.

[18] Wani JA, Sharma S, Muzamil M, Ahmed S, Sharma S, Singh S. Machine learning and deep learning based computational techniques in automatic agricultural diseases detection: Methodologies, applications, and challenges. Archives of Computational methods in Engineering. 2022 Jan;29(1):641-77.

[19] Binson VA, Manju G. Automated Disease Detection in Silkworms Using Machine Learning Techniques. Advance Sustainable Science Engineering and Technology. 2024 Sep 20;6(4):02404015

[20] Janssens AC, Martens FK. Reflection on modern methods: Revisiting the area under the ROC Curve. International journal of epidemiology. 2020 Aug;49(4):1397-403.