# Sparse Modelling for Feature Learning in High Dimensional Data

Harish Neelam

Michigan State University  
CMSE Department   
+1 (517)-507-7176

[neelamha@msu.edu](mailto:neelamha@msu.edu)

Koushik Sai Veerella

Michigan State University  
CMSE Department   
+1 (469)-601-1149

[veerella@msu.edu](mailto:veerella@msu.edu)

Souradip Biswas

Michigan State University  
CMSE Department   
+1 (517)-944-9381

[biswasso@msu.edu](mailto:biswasso@msu.edu)

**ABSTRACT**

This project introduces a robust pipeline for wood surface defect detection employing various machine learning techniques. The dataset comprises images of wood surfaces, and the primary objective is to process, train models, and assess their performance in defect detection. The workflow encompasses data loading, preprocessing, feature extraction and learning, dimensionality reduction, and model training. Notably, the code leverages pre-trained models like VGG19 and incorporates anomaly detection methods such as Isolation Forest and Local Outlier Factor. Evaluation metrics, including accuracy, F1 score, and visualizations, are employed to gauge model performance. The project's distinctive focus lies in addressing the challenge of extracting meaningful features from high-dimensional datasets. Sparse modeling techniques, specifically Lasso and proximal gradient methods, are employed to efficiently reduce dimensionality while enhancing interpretability. The overarching goal is to advance the field of wood surface defect detection by providing a comprehensive framework that combines state-of-the-art machine learning methods with sparse modeling techniques, ultimately contributing to more accurate and interpretable results.

**Keywords**

Wood defect detection, Feature Reduction, Sparse Modelling, Lasso, Elastic Net, Proximal Gradient Descent, Imbalance Dataset, KNN, Binary Classification.

# INTRODUCTION

Wood surface defect detection is a critical aspect of quality control in various industries, ensuring the production of high-quality wooden products. In recent years, the integration of machine learning techniques has shown promising results in automating this process.

Our project aims to contribute to incorporate a variety of machine learning techniques, such as pre-trained models (e.g., VGG19) and classification methods, to leverage different approaches for identifying wood surface defects. Also, to focus on addressing the challenges posed by high-dimensional datasets by incorporating sparse modelling techniques, specifically Lasso and proximal gradient methods, to efficiently extract meaningful features while reducing dimensionality. And to emphasize the importance of interpretability by utilizing sparse modelling techniques that not only reduce dimensionality but also provide insights into the features contributing to defect detection. This aims to make the results more understandable and actionable. It also aims to make the project outcomes practically relevant for industries involved in wood processing and manufacturing by demonstrating how the improved defect detection capabilities can lead to tangible benefits such as cost savings and quality improvement.

Dimensionality reduction through sparse modelling contributes to improved computational efficiency during both training and inference. In sparse models like Lasso, the non-zero coefficients directly indicate the importance of corresponding features. The sparse nature of the models results in a set of human-interpretable features that are directly linked to the presence or absence of defects. This interpretability is vital in scenarios where the model's decisions need to be explained to non-experts or integrated into decision-making processes.

The incorporation of diverse machine learning techniques, including pre-trained models and classification methods, significantly improves the accuracy of identifying and classifying wood surface defects. This enhancement is crucial for industries involved in woodworking and quality control, ensuring the production of high-quality wooden products. By demonstrating tangible benefits such as cost savings and quality improvement, the project makes its outcomes practically relevant for industries engaged in wood processing and manufacturing.

The focus on interpretability through sparse modelling techniques provides actionable insights into the characteristics of wood surface defects. This transparency allows stakeholders and decision-makers to understand the features contributing to defect detection, facilitating informed decision-making processes. Sparse modelling promotes the selection of a concise set of informative features, mitigating over-reliance on specific attributes. This project's contributions extend beyond wood surface defect detection, impacting both the wood industry and the field of feature optimization by leveraging machine learning techniques and emphasizing the importance of interpretability and efficiency.

# RELATED WORK

Feature selection process is a crucial preliminary step in handling high-dimensional datasets. This process aims to reduce dimensionality by selecting a subset of features that effectively capture the distinctions among features concerning the type of label.

[2] R. Muthukrishnan and R. Rohini in their published paper, LASSO: A feature selection technique in predictive modeling for machine learning, explored the features of the popular regression methods, OLS regression, ridge regression and the LASSO regression. The performance of these procedures has been studied in terms of model fitting and prediction accuracy using real data and observed promising results.

[6] Maryam A. Alghamdi, Mohammad Ali Alghamdi, Naseer Shahzad, Hong-Kun Xu discussed regarding the iterative methods for solving the lasso which include the proximal-gradient algorithm and the projection-gradient algorithm in their article, Properties, and Iterative Methods for the Lasso.

In our project, we recognize the significance of addressing these challenges in feature selection, especially in the context of wood surface defect detection. We aim to explore and potentially extend existing methods, incorporating approaches that consider correlated features while balancing computational efficiency and addressing class imbalance concerns. Our focus is on developing a feature selection strategy that aligns with the unique characteristics of our dataset and enhances the interpretability and performance of our defect detection models.

# DATA ANALYSIS

The dataset we used in our project was Wood Defection dataset [1]. It contains 4000 images with annotation for wood surface defects of different types. Some of the examples include No Defect and defect (Quartzity, Live knot, Marrow resin, Dead knot, knot with crack, knot missing and Crack). The original dataset had high resolution images captured with special camera which takes up to 12MB of disk space per image. The size of the dataset has been reduced by resizing the images to 256\*256\*3 (196608 features). The dataset contains YOLOv5 annotations, which contains the bounding boxes and respective labels.

The task is framed as a binary classification problem with the objective of detecting defects in wood surfaces. The dataset is structured such that the presence or absence of a defect is indicated by the contents of a bounding box associated with each image. A bounding box devoid of any labels denotes Class 0, signifying no defect, while a bounding box containing any label designates Class 1, indicating a defective wood surface. A notable challenge is the imbalanced distribution of data, with Class 1 instances being approximately one-tenth the number of Class 0 instances. This data imbalance can lead to biased models, particularly favoring the majority class. To address this, the code employs various sampling techniques during training to balance the representation of both classes. To effectively address the imbalanced nature of our dataset, we employed the Random Over-Sampling technique.

# METHODOLOGY

Pretrained models, trained on massive datasets, have demonstrated exceptional capabilities in learning hierarchical representations of features from raw data [7]. This section introduces the concept of feature representation and its significance in various computer vision applications, and how these features are used to select important features which influence the outputs.

## Feature Representation

For this wood surface defect detection project, we leveraged powerful pre-trained models called ResNet50(Residual Network), AlexNet, VGG19 and EfficientnetB7. By selecting VGG19, we aimed to harness its superior ability to capture intricate details and patterns from wood surface images, facilitating effective defect detection. It emerged as the top performer and we chose it as the preferred model for our application.

Due to its simplicity and popularity, VGG19 is often used for transfer learning. Pre-trained models on large datasets like ImageNet are available, and these models can be fine-tuned for specific tasks with smaller datasets. This was particularly useful for our specific problem. The features extracted by VGG19 act as a form of representation of our data that may be more informative than raw pixel values, especially for image-related tasks. With this model we were able to generate 32768 features for each image as shown in Fig. Proceeding with this, we built a baseline model with Random Forest and KNN using the Imbalanced data and Oversampled data.

## KNN Classification Model

K-Nearest Neighbors (KNN) is a simple yet powerful algorithm in machine learning that falls under the category of supervised learning. It is primarily used for classification and regression tasks. The fundamental idea behind KNN is to classify a data point based on the majority class of its k nearest neighbors in the feature space. The choice of the parameter k influences the model's sensitivity to local patterns, with smaller values leading to more flexible and potentially noisy predictions, while larger values may result in a smoother decision boundary.Its performance can be sensitive to the choice of distance metric and the appropriate value of k. The KNN model was trained and tested accordingly, yielding the following metrics. KNN with neighbors set to 1 was employed; when increased, the correctly classified Class 0 samples tended to shift towards Class 1. Therefore, a choice of n\_neighbors = 1 was maintained for all subsequent models.

* Imbalanced Data:
  + Accuracy: 84.875%
  + F1-Score: 0.8376
* Oversampled Data
  + Accuracy: 88.49%
  + F1-Score: 0.892

## Feature Reduction

PCA, a linear dimensionality reduction technique, systematically identified and retained the most critical features by projecting the data onto a set of orthogonal axes, known as principal components. By exploring various combinations of these components, we sought to strike an optimal balance between reducing the computational complexity of the model and preserving the essential information required for accurate defect identification. The highest accuracy attained with PCA configurations highlighted its limited effectiveness in improving our model's performance for the given dataset.

* Accuracy: 85.35%
* F-1 score: 0.8594.

Complementing the efforts of PCA, we employed Kernel Principal Component Analysis (Kernel PCA) to capture non-linear structures within our wood surface image data. Kernel PCA extended the capabilities of traditional PCA by leveraging kernel functions to map the data into a higher-dimensional space. This facilitated the extraction of complex patterns that might not be discernible in the original feature space. By systematically testing different combinations of kernel parameters and components, we aimed to identify the most effective configuration for our defect detection model. The highest accuracy achieved through Kernel PCA underscored its limited effectiveness in revealing intricate relationships within the data, posing challenges to the success of our wood surface defect detection system.

* Accuracy: 85.96%
* F-1 score: 0.8646.

There could be many reasons why PCA and Kernel PCA didn't perform well. They are:

**Loss of Information:** PCA and Kernel PCA involve dimensionality reduction, potentially leading to a loss of crucial information in the data and impacting the accuracy.

**Overfitting Concerns:** Dimensionality reduction techniques, including PCA and Kernel PCA, may introduce overfitting risks.

**Dense Solutions:** PCA and Kernel PCA typically provide dense solutions by considering all features in the transformed space thus impacting the model's performance.

# SPARSE MODELLING

In our continuous quest to enhance the performance of our wood surface defect detection project, we explored feature learning techniques with a specific focus on sparse modeling, a pivotal aspect of our project goals. Contrary to the expected improvement, feature reduction techniques did not yield the desired enhancements in performance. As a strategic pivot, we shifted our attention to sparse modeling techniques, aiming to selectively emphasize relevant features by inducing sparsity in the feature space. Two prominent sparse modeling techniques were implemented: Lasso regularization and Elastic Net regularization. The rationale behind employing sparse modeling was to accentuate the significance of relevant features while mitigating the impact of irrelevant ones. [2] This strategic shift toward sparse modeling reflects our commitment to adapt and refine our approach based on empirical results, ultimately steering the project toward its primary goal of efficient and accurate defect detection in wood surfaces.

## Lasso Regularization

Lasso regularization serves as a potent tool in preventing overfitting, a common challenge in machine learning models. Its mechanism involves augmenting the standard least square’s objective function with a penalty term proportional to the absolute values of the coefficients within the regression model as shown in the Equation (1). This added L1 penalty induces sparsity in the model by driving certain coefficients to exactly zero. In the context of our wood surface defect detection project, we applied Lasso regularization to optimize the cost function, which includes both the least square errors and the L1 penalty term. During the training phase, where the model learns from labeled data, the objective is to minimize this cost function. [4] This process entails determining the coefficients that minimize the combined impact of least square errors and the L1 penalty. These coefficients, once identified through the training process, are then employed to predict features in testing instances. The regularization parameter, denoted as lambda, plays a pivotal role in governing the strength of regularization. A higher value of lambda indicates more potent regularization, effectively emphasizing sparsity in the model. This fine-tuning parameter allows us to strike a balance between fitting the model to the training data and preventing it from becoming overly complex, thereby contributing to robust and effective defect detection in wood surfaces.

--- *Eq (1)*

The Lasso regularization was implemented using the library from scikit-learn. Setting the regularization constant (Lambda) to 0.01, the application of Lasso regularization resulted in a notable reduction of features for each image. Specifically, the feature count diminished from 32,768 to 2,026, signifying a substantial simplification of the dataset. The efficacy of this reduction lay in its ability to retain relevant information while discarding less significant features. The refined set of 2,026 features was then employed to train and fit a KNN model. This sequential approach not only facilitated computational efficiency but also aimed to enhance the model's performance by focusing on the most informative attributes derived through Lasso regularization.

* Accuracy: 89.12%
* F-1 score: 0.8974.

Elastic Net regularization [8], a hybrid approach encompassing both L1 and L2 regularization, offers a nuanced solution by striking a balance between the sparsity-inducing nature of Lasso and the grouping effect of Ridge as shown in equation (2). In our implementation, we set the hyperparameters to an alpha value of 0.01 and an L1\_ratio of 0.5, signifying equal importance given to both Lasso and Ridge regularization. Notably, a L1\_ratio of 1 emphasizes Lasso, while 0 emphasizes Ridge, allowing us to fine-tune the regularization strategy based on the characteristics of our data. Remarkably, our experimentation with Elastic Net yielded results comparable to those obtained with Lasso in terms of selected features and performance scores. However, the similarity in results suggests that the functionality of Ridge, a prominent component of Elastic Net, might not be particularly suitable for our specific case. Ridge regression is particularly beneficial when dealing with datasets that exhibit highly correlated features.The adaptive nature of our approach acknowledges the nuances of our dataset, underscoring the importance of tailored regularization strategies to optimize the performance of our wood surface defect detection model.

-- *Eq (2)*

## Proximal Gradient Descent

With the streamlined dataset obtained through Lasso regularization, the focus shifted towards optimization strategies to further enhance the model's efficiency and predictive accuracy. To further refine the feature representation and enhance the overall model performance, the proximal gradient optimization technique was incorporated. This approach, extending beyond traditional optimization strategies, allowed for fine-tuning the features derived through Lasso regularization. By leveraging proximal gradient methods, the model aimed to strike a balance between sparsity and accuracy, encouraging the retention of critical information while minimizing irrelevant details.

We applied the proximal gradient descent method to the Lasso features to further optimize Lasso features and other penalties (L1, L2, Elastic net, Lasso Group using the following objective function equation (3):

*Objective Function* = 0.5 \* ||Ax – b||22 + λ ||x||1

---*Eq (3)*

A represents the Lasso features, x signifies the coefficients, initialized with zeros, and b corresponds to the training target values. The regularization parameter λ is set to 0.1, indicating the strength of the regularization. The choice of the L-BFGS-B method during this optimization process is significant. To initiate the optimization, the coefficients x starts with zeros, and the process iterates for 10 cycles. The utilization of the 'L-BFGS-B’ method within the SciPy library facilitates the efficient minimization of the defined objective function, ultimately refining the coefficients for enhanced performance in defect detection.

The results of the optimization process are truly noteworthy, culminating in a refined feature set comprising only 1,179 features. This substantial reduction from an initial pool of 32,768 features signifies a highly impactful dimensionality reduction, streamlining the model and enhancing its interpretability. The most striking aspect is that, despite the relatively small number of features, the wood surface defect detection model achieved outstanding performance metrics.

* Accuracy: 89.65%
  + F-1 score: 0.902.

# EVALUATION

This section discusses the advantages and comparison of feature reduction and sparse modelling Techniques.

## Advantages of Lasso and PGD

The integration of Lasso regularization and proximal gradient optimization offers several distinct advantages in the context of feature enhancement and model refinement. By penalizing the absolute values of coefficients, these techniques actively promote sparsity, potentially driving certain coefficients to zero and effectively eliminating non-essential features from the model.

Moreover, these methods prove to be highly advantageous in the realm of high-dimensional datasets, where feature selection is paramount for both model interpretability and performance. Lasso stands out for its feature selection capabilities, shedding light on the most pertinent features and rendering the model more interpretable. The combination of Lasso and proximal gradient ensures a streamlined and computationally efficient modeling process.

## Sparse Modelling vs Feature Reduction

PCA, while effectively reducing the number of features, is fundamentally indifferent to class labels. Its primary objective is to retain maximum variance in the data, making it less tailored for classification tasks. In contrast, L1 regularization, as exemplified by Lasso Regression, prioritizes features that exhibit high correlation with class labels. By driving certain features toward zero, L1 regularization inherently aims to decrease the number of features while emphasizing those with significant predictive power.

PCA is advantageous when variables contribute equally to variance or when there is high multicollinearity among features. In contrast, Lasso Regression is a preferable choice when variable elimination is possible, and the dataset has been cleansed of multicollinearity.

# CONCLUSIONS

The success of Lasso regularization and Proximal Gradient Descent (PGD) underscores the importance of tailoring feature selection strategies to the characteristics and dimensionality of the dataset at hand. Despite the limited number of features, both Lasso and PGD exhibited commendable performance, showcasing their ability to capture essential information and facilitate accurate classification. The judicious application of these sparse techniques not only streamlined the feature space but also contributed to the overall robustness and interpretability of the model. The results emphasize the significance of choosing techniques that align with the inherent structure of the data, particularly when faced with challenges such as dimensionality reduction and feature relevance.

# REFERENCES

1. *Kodytek Pavel, Bodzas Alexandra, & Bilik Petr. (2021). Supporting data for Deep Learning and Machine Vision based approaches for automated wood defect detection and quality control. [Data set]. Zenodo. https://doi.org/10.5281/zenodo.4694695*
2. *R. Muthukrishnan and R. Rohini, "LASSO: A feature selection technique in predictive modeling for machine learning," 2016 IEEE International Conference on Advances in Computer Applications (ICACA), Coimbatore, India, 2016, pp. 18-20, doi: 10.1109/ICACA.2016.7887916.*
3. *Ian T Jolliffe, Nickolay T Trendafilov & Mudassir Uddin (2003) A Modified Principal Component Technique Based on the LASSO, Journal of Computational and Graphical Statistics, 12:3, 531-547, DOI:*[*10.1198/1061860032148*](https://doi.org/10.1198/1061860032148)
4. *C. Park, “Simple principal component analysis using Lasso,” Journal of the Korean Data and Information Science Society, vol. 24, no. 3. Korean Data and Information Science Society, pp. 533–541, 31-May-2013.*
5. *A. Argyriou, T. Evgeniou, and M. Pontil. Multi-task feature learning. In Proc. of NIPS, 2006.*
6. *Maryam A. Alghamdi, Mohammad Ali Alghamdi, Naseer Shahzad, Hong-Kun Xu, "Properties and Iterative Methods for the Lasso", Abstract and Applied Analysis, vol. 2013, Article ID 250943, 8 pages, 2013.* [*https://doi.org/10.1155/2013/250943*](https://doi.org/10.1155/2013/250943)
7. *Hassan, Aya Gamal Nasr et al. “Image classification based deep learning: A Review.” Aswan University Journal of Sciences and Technology (2022): n. pag.*
8. *H. Zou and T. Hastie. Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67(2):301–320, 2005.*