Revolutionizing Boiler Maintenance: Image Processing Techniques for Boiler Scale Detection

BY

Md Shohidul Islam Polash
ID: 241-25-042

This Report Presented in Partial Fulfillment of the Requirements for
The Degree of Masters of Science in Computer Science and Engineering

Supervised By

Dr. S. M. Aminul Haque
Professor
Department of CSE
Daffodil International University

Co-Supervised By

Dr. Abdus Sattar
Associate Professor
Department of CSE
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

MAY 2025

APPROVAL

This Project/Thesis titled **"Revolutionizing Boiler Maintenance: Image Processing Techniques for Boiler Scale Detection"**, submitted by **Md Shohidul Islam Polash,** ID No: **241-25-042** to the Department of Computer Science and Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of **MSc. in Computer Science and Engineering** and approved as to its style and contents. The presentation has been held on **24-05-2025.**

BOARD OF EXAMINERS

Chairman

Prof. Dr. Sheak Rashed Haider Noori
Professor and Head
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner

Dr. Md. Zahid Hasan
Associate Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner

Dr. Naznin Sultana
Associate Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

External Examiner

Dr. Md. Zulfiker Mahmud
Professor
Department of Computer Science and Engineering
Jagannath University

DECLARATION

I hereby declare that this research has been done by me under the supervision of **Dr. S. M. Aminul Haque, Professor, Department of CSE,** Daffodil International University. I also declare that neither this project nor any part of this project has been submitted elsewhere for award of any degree or diploma.

Supervised by:

Dr. S. M. Aminul Haque
Professor
Department of CSE
Daffodil International University

Co-Supervised by:

Dr. Abdus Sattar
Associate Professor
Department of CSE
Daffodil International University

Submitted by:

Md Shohidul Islam Polash
ID: 241-25-042
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

First I express my heartiest thanks and gratefulness to Almighty Allah for His divine blessing which makes it possible to complete the final year project/internship successfully.

I am really grateful and wish my profound indebtedness to **Dr. S. M. Aminul Haque, Professor, Department of CSE,** Daffodil International University, Dhaka, deep knowledge & keen interest of my supervisor in the field of Machine Learning to carry out this project. His endless patience, scholarly guidance ,continual encouragement , constant and energetic supervision, constructive criticism , valuable advice ,reading many inferior drafts and correcting them at all stages have made it possible to complete this project.

I would like to express my heartiest gratitude to **Dr. Sheak Rashed Haider Noori**, **Head,** Department of CSE, for his kind help to finish our project and also to other faculty members and the staff of CSE department of Daffodil International University.
I gratefully acknowledge the valuable data and insights provided by various industries in Bangladesh, whose cooperation significantly contributed to the successful execution of this project. In particular, I am thankful to **Md. Sharafat Ali, Deputy Chief Inspector of Boilers, Ministry of Industry, Bangladesh**, for his generous assistance and cooperation during the data leveling process.

Finally, I must acknowledge with due respect the constant support and patients of my parents.

ABSTRACT

Scale formation is a problem that affects the performance of industrial process systems, reducing thermal efficiency, consuming extra energy, and causing equipment breakdown. But so far, although identification and classification of scale type are also of vital importance in industry, there is still no automatic scale identification system available to scale types from a visual perspective. However, the automatic identification of organic scale on boiler has not been studied in the prior research. In this paper, the first such system is proposed based on machine learning and deep learning to classify boiler scale deposits from the real industrial images that, belonging to the most frequent types (CaCO3, Fe3O4, Miscellaneous Scales), discriminated so far. A real world dataset was created with the images obtained directly from the operation of the boiler systems of the industries in Bangladesh and were annotated by the experts of the domain. To improve the model accuracy and robustness, we tested different pre-processing pipelines and feature extraction methods including classical descriptors (HOG, LBP, GLCM), CNN embeddings (VGG16, 2DSCN), and hybrid methods. Between the classical machine learning classifiers the optimized XGBoost attained the highest accuracy of 90.12%. In the realm of deep learning, we introduce a ScaleNet V1 with custom-designed blocks inspired by both deep residual networks and squeeze-and-excitation modules, which also includes a trainable attention module and is designed in Keras Tuner. The results showed that ourScaleNet V1outperformed conventional networks such as EfficientNet V2L,, MobileNet V2 and ResNet18, by achieving a test accuracy of 93.42%, lower prediction latency, and higher class-wise results. We validated the contribution of each architectural component with an ablation study and performed LIME-based explainability during the training phase of the model, gaining interpretability of model decisions, and improving the industrial applicability and trust. The proposed framework provides for the first time a scalable, interpretable, efficient and real-time solution for the boiler scale monitoring, which can constitute another benchmark in intelligent maintenance systems in industrial process optimization.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1
INTRODUCTION

1.1 Introduction

Boilers are employed in various industries such as heating systems, power systems and their processes and many other industries that can’t possibly function without using heat, such as manufacturing or processing, depends upon the thermal power to perform mechanical work or for chemical processing, hence they are essential for industries. Smelly water, loud noises and leaking from bursting hoses are some of the ongoing problems but one constant issue is scale deposits on internal surfaces from dissolved impurities in feed water. The deposits usually contain calcium carbonate, long with iron oxides and other mineralogical elements, causing a great decrease in heat transfer efficiency, fastening of fuel consumption, corrosion and, in severe cases, a damage leading to a boiler breakdown. Accordingly, prediction and correct classification of boiler scale types is vital especially in predictive maintenance and industrial safety and efficiency [1], [2].

Generally, the scale detection method relies on the experience of a probe's technology and component analysis, which is a time consuming and labor-intensive process and not suitable for large-scale industrial field monitoring on-line [1]. Machine learning (ML) and deep learning (DL) have the potential to automate visual inspection in several industrial fields [1], [2], [3]. Nevertheless, the current approaches remain generally applied to the basic detection of defects and encounter difficulties in the classification of boiler scales on account of the large number of visual parallels and texture liplaps among various scales. This limitation highlights the necessity of using a dedicated framework for boiler scale classification.

To address these limitations, our study introduces an interpretable, robust and scalable deep learning based system to automatically classify the fouling content of industrial-scale boiler images. This work is also done in collaboration with industrial partners in Bangladesh which is the first time of its kind in the area of sustainable shipping and we have been able to build a database through these partners through their modern fleet data for Bangladesh. This data captures the real-world diversity through three major types:

Calcium Carbonate Scale, Iron Oxide Scale, and Miscellaneous Scale Types. The paper systematically investigates various preprocessing pipelines and deep feature extraction methods, as well as diverse machine learning classifiers. We also propose a new custom architecture called ScaleNet V1 (also custom) which is tuned with Keras Tuner technique and finally it is built on top of various state of the art modules such as the residual connections and the attention modules and the squeeze-and-excitation networks to name a few.

The proposed technique demonstrates encouraging potential and is verified with extensive experimental results. ScaleNet V1 could change the game in real deployment of boiler applications with better classification accuracy, generalization, and inference time, thus a bright future of industrial maintenance.

The main contributions of the proposed work are as follows:

- Novel Dataset for Boiler Scale Classification: We present a newly collected and carefully annotated dataset of industrial boiler scale images corresponding to real world factory environments in Bangladesh and divided into Calcium Carbonate, Iron Oxide, and Miscellaneous scales. This dataset targets a novel aspect of computer vision and industrial maintenance.
- Custom Deep Network - ScaleNet V1: We introduce ScaleNet V1, a new hybrid CNN architecture tailored for high accuracy scale classification. The model is built up by residual blocks, a developed attention mechanism and SE modules, which is capable of extracting both local texture and global context that are important for distinguishing visually similar scale types. To thoroughly evaluate, we benchmarked on several state-of-the-art CNN models (EfficientNet V2L, MobileNet V2, ResNet18) and compare with them and show the superiority of our ScaleNet V1 w.r.t test accuracy (93.42%), prediction time and class-level robustness.
- Extensive Preprocessing and Augmentation: Extensive testing of various image enhancement and augmentation approaches (e.g., CLAHE, contrast scaling, gamma correction) resulted in establishing the most effective preprocessing pipeline for generalization and robust model training.

- Feature Extraction Benchmarking: We compare several conventional as well as deep feature extraction methods including 2DSCN + SIFT, LBP, GLCM, and pretrained VGG16 with a combination of local descriptors and contrastive learning approach achieved the best performance in ML.
- ML Model Selection and Tuning: From 11 traditional ML classification models, we selected and optimized a lightweight and high-performance ML model (Tuned XGBoost) with a Randomized Cross-Validation over 900 parameter settings, striking a balance between accuracy and inference efficiency.
- Explainability using LIME: To make our method transparent, reliable and safe for industrial settings, we included LIME (Local Interpretable Model-Agnostic Explanations), through which domain experts can trace which image regions led to the decision made by the model. This is sugar for trust, and fuel for performant diagnostics.

This paper advances the state-of-the-art on intelligent maintenance systems by bridging the gap between computer vision and a domain-specific industrial problem, showing a reliable and explainable approach for boiler scale classification automation.

1.2 Motivation

Industrial boilers are a critical part of the fabric of many manufacturing and energy sectors today, and the size of the loss for even a fraction of inefficient operation can be highly detrimental to the economy and the industry. If there is a menace to boiler efficiency, one that is arguably most pervasive is the build-up of scale deposits, especially of calcium carbonate, iron oxide, and minerals. These deposits hinder heat transfer, increasing fuel use and material degradation. The automobile scale types are identified and classified through manual or chemical methods, which are time-consuming, inconsistent, and far from adaptable to the modern industrial scalability demand.

The invention of computer vision and deep learning realistically can provide a paradigm shift in automating this classification process with accuracy and efficiency. However, image-based boiler scale classification has not been studied yet, creating a significant lack in the existing intelligent maintenance systems. Motivated by the need for accurate, real-time, and interpretable AI-based solutions, this research will enable industries to make decisions ahead of time to avoid downtimes and increase operational efficiency.

1.3 Rationale of the Study

There has been no previous automated work on visual boiler scale classification, validating the need for the current study due to the importance of the task to industrial safety and costs. Existing inspection activities are human error-prone, relatively small, and more reactive than preventive. We introduce a deep learning-based classification framework trained on a novel real-world dataset to fill the gaps between industrial inspection needs and AI-driven automation.

In addition, ScaleNet V1 is a base model with attention-based probes, residual connections, and squeeze-and-excitation modules, designed as a customized architecture for this particular classification challenge. We also use LIME to enhance explainable AI integration in a manner that promotes its practical adoption via interpretation and trust in the model predictions by industrial stakeholders. Therefore, it is significant for practice, academia, and technology to improve intelligent inspection systems in the industry.

1.4 Research Questions

The investigation should be headed and based on where contributions are focused, and the study intends to answer the following important research questions:

RQ1: Can we create a deep learning model to correctly classify boiler scale images into their classes (Calcium Carbonate, Iron Oxide, Miscellaneous)?

RQ2: What are the best image preprocessing and feature extraction methods for obtaining scale classification accuracy?

RQ3: ScaleNet V1 model vs. existing state-of-the-art ML & DL models: What are the improvements in the parameters of interest (e.g., generalizability, accuracy, efficiency) of the proposed ScaleNet V1 model with respect to existing methods?

RQ4: Are interpretability methods such as LIME used to explain model predictions in an industrial setting?

RQ5: What are their feasible compromises regarding model performance, prediction time, and the ability to classify in practice?

1.5 Expected Output

This research is expected to produce the following outcomes:

- An image dataset of boiler scale images categorized by experts in their field.
- ScaleNet V1 a deep learning backbone that has been customized for classification tasks like this one and out-performed general CNN architectures
- Selection of optimal preprocessing and feature extraction methods for boiler scale images
- Full benchmarking of standard ML and deep learning models with respect to accuracy, prediction time, and class/target-wise benchmarking.
- Implementation of an explainable AI framework with LIME, which gives the user a visual understanding of the model decision-making for real-world application
- ScaleNet V1 ablation study for analyzing performances of architectural components
- An automated classification system that is reproducible, efficient, scalable, and can be deployed in industrial maintenance pipelines for real-time decisions.

1.6 Project Management and Finance

The research work doesn't get fund from any individuals or organization.

1.7 Report Layout

Chapter 1 sets out the study's background, aims, and main research questions. In Chapter 2, brief summaries of the literature review are presented. We explain the method thoroughly in Chapter 3. In Chapter 4, we present the experimental results of the paper. Chapter 5 presents the affordance plan and the social and environmental impacts, as well as discusses ethical aspects. Chapter six's last chapter summarizes the present study and proposed approach for further studies.

CHAPTER 2
BACKGROUND

2.1 Boiler Scale and Its Industrial Effects

Boiler scale, originating mainly from mineral impurities in feedwater, is still a major issue in industrial thermal equipment. Typical types are calcium carbonate, iron oxide, chromium oxide and silica-type compounds. These deposits cause decrease in heat transfer and increase in operation costs which results in degradation and failure of equipment [4], [5]. Conventional detection approaches, for example visual detection analysis, XRD and SEM, are typically time-consuming, labor-intensive and do not provide real time or scalable solutions [5].

2.2 The Requirement for An Intelligent Boiler Inspection

Manual and chemical detection methods are commonly utilized, but are subjective, have a slow turnaround, and can cause downtime when not properly performed. Recent study highlights the need for intelligent, autonomous classification systems for predictive maintenance, efficiency improvement, and scaling in large-scale industrial infrastructures [6], [7], [8]. Yet, no prior research has addressed the machine/deep learning-based classification of boiler scale deposit to our knowledge, therefore this work is pioneering in this field.

2.3 Relate work in Machine Vision for Industrial Defect Detection

In industrial sectors (e.g., corrosion detection, crack identification, weld inspection, etc.), machine vision systems have been extensively utilized. Early systems were based on traditional image processing techniques) (rule-based thresholding, statistical filters, handcrafted features (e.g., LBP, GLCM)) for detection of the surface anomaly [8], [9]. Such methods, although fast and interpretable, are not adaptive to subtle, complex, or erratic defects.

With the emergence of deep learning, CNN structure models such as ResNet, VGG16, EfficientNet and YOLO have shown great capability, generalization and real-time in defect classification tasks [10], [11], [12].

2.4 Machine Learning and Feature Engineering for Industrial Image Classification

If combined with strong pre-processing and feature extraction mechanisms, classifiers like Random Forest, SVM, XGBoost are widely used for surface level classification [13],[14]. Common pre-processing measures, for instance, contrast stretching, histogram equalization and bilateral filtering are used for enhancing the quality and consistency of the images. Texture-based features like GLCM, LBP or SIFT and HOG features are also powerful in practice, especially when used in conjunction with ensemble classifiers or deep features [15], [16].

2.5 Deep Learning Models and Improvements

Industrial classification tasks have been revolutionized by deep learning. Classical architectures are also used in some research for detecting surface defects, such as VGG16, ResNet, EfficientNet, and Inception, which can hierarchically learn the features [17], [18]. Recent works have incorporated attention mechanisms, SE blocks and residual connections, in order to further boost performance and interpretability of the model, and to better focus on task-related features and to avoid the vanishing gradient [19], [20], [21]. Various methods have been proposed for developing domain-specific custom CNN architectures or hybrid models for domain-specific tasks that significantly outperform standard networks since they have stronger context sensitivity and modular structure [22], [23], [24].

2.6 Data Complexity in Industrial Classification

Datasets in industrial settings Face class imbalance, low sample size andThere is uneven quality in the images. Although, there are many publicly available datasets for types of defects such as corrosion, steel surface defects, spot weld faults etc., no standard dataset exists for boiler scale classification. This again underscores the novelty and importance of the datasets, like the one proposed in this study [25], [26], [27]. Approaches such as data augmentation, GAN-based synthesized data, cost-sensitive learning are widely used for addressing such problem [28], [29], [30].

2.7 Metrics and Explainability in Industrial Models

Accuracy, precision, recall, and F1-score are still the traditional evaluation measures that are used for comparison of the performance of various classifiers. Nevertheless, the deployment in the context of the real-time scenario requires both computational tradeoffs and smallish models [31], [32]. Techniques such as LIME and Grad-CAM as well as other attention-based visualization methods are introduced to aid model interpretation, toward industrial use and compliance with regulations [33].

2.8 Perceived Gaps and Direction of Research

Even though there exists a plenty of literature on industrial inspection and defect classification, to the best of our knowledge, the classification of boiler scale has not been considered with ML or DL. The bottleneck of the domain lies in its challenging issues: overlapping visual content, inconsistent temporal patterns, and absence of dataset annotation [34], [35], [36]. This gap is bridged by this study, where:

- A **new real-world dataset**,
- A **custom hybrid CNN architecture** and
- The **integration of explainability** tools for practical deployment.

The research not only contributes technical advancements but also establishes a **new subdomain** in intelligent industrial maintenance systems.

CHAPTER 3

RESEARCH METHOLOGY

3.1 Proposed Methodology/Applied Mechanism

This methodology focuses on detecting boiler scale using advanced machine learning and deep learning techniques. It ensures accurate and explainable results to support effective decision-making in industrial settings.

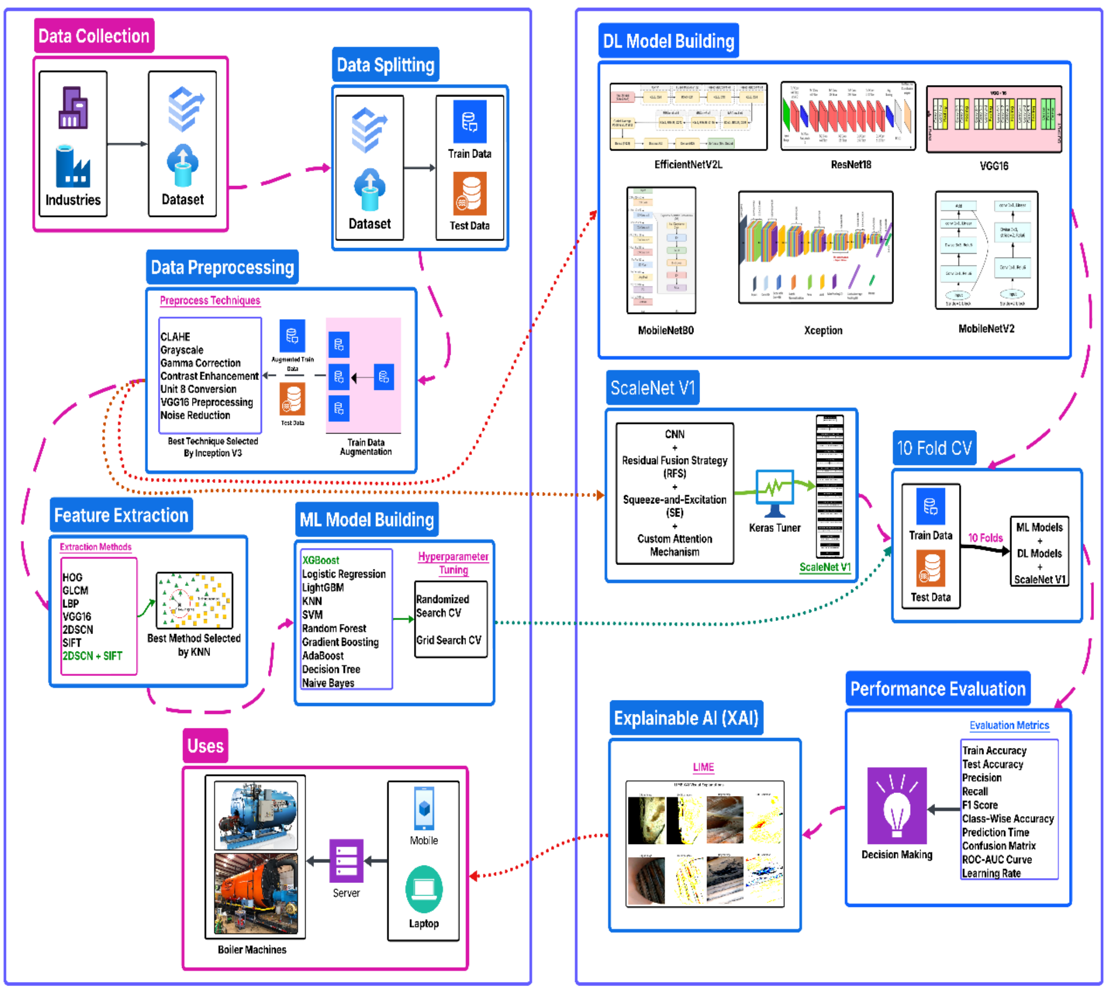


Fig 3.1: The working process to perform Brain disease prediction

3.2 Data Collection

The focus of this research, the boiler scale industrial image data, was directly gathered from various factories in Bangladesh. It was intended to discover and classify the most common fashions of scale deposition of industrial boilers on account of the particular water chemistry of the country. The dataset was categorized by Ministry of Industry experts in Bangladesh. The dataset was annotated divided into three main categories as follows:

Calcium Carbonate Scale: 415 images(class having the highest number of scale) This is because the proportion of calcium and scale in the water in Bangladesh is going up but the causes of scaling in industrial boilers have increased mainly because the water system of the country is no longer the same as it used to be.

Scale Pictures: This Album (269) Iron Oxide Scale: This Album (6 pictures) Idiopathic boiler scales are less common than calcium carbonate scales and iron oxide scales are usually the result of corrosive conditions in the boiler itself.

Others: 278 images; This class considers rare scales of formations, e.g., Silica Scale, Magnesium Silicate Scale, and etc. Since they have a low number of samples each, we put them all together and designated them as "Others" so as that they could be equally considered for classification without favor the model.

A visual breakdown of the dataset distribution is shown in Figure 3.2.

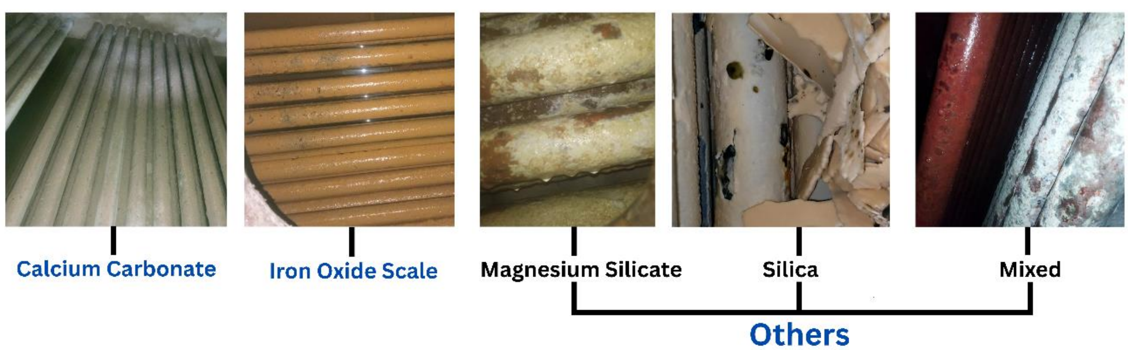


Fig. 3.2: Images of Different Types of Scales from Boiler Machines

3.3 Data Preparation

We conducted certain necessary preprocessing on the collected dataset so that the model runs at its peak and generates correct results. Images were resized and normalized to the same input size and pixel intensity. Same for the dataset, it was shuffled to avoid any affinity bias in training the model.

3.3.1 Train-Test Split

For the limited number of records, the data were divided into training and testing parts with a common 75%-25% proportion to verify the performance of the model. This approach fishes out the most from the data and also checks its accuracy on a chunk of the data that it has not seen before (that provides a good approximation of its generalization ability)`. These included 719 for training and 243 for testing. This 50/50 division is arbitrary in the sense that class balancing was preserved, and the representativeness of each type in the scale was maintained after splitting. The split was done immediately after data acquisition and prior to any data preprocessing and model selection, to avoid leakage or bias. The splits were also tested (eg. 80%-20%, 70%-30%) with the base model but we chose to use the 75%-25% split for all the subsequent experiments since not only they gave the most accurate answer but also demonstrated that the parameters generalized well.

3.3.2 Data Augmentation

As the dataset is small, the model can overfit to the few examples in the dataset so in optimization scenarios with respect to those with some data augmentation to generalize better. This procedure artificially modifies and augments the training set by randomly transforming the input images.

The requirement of augmentation in this work is essential, since industrial image samples will tend to contain significant variability in illumination condition, orientation and position. This makes the model robust to real world boiler variations as these are incorporated during the training.

The augmentation pipeline was applied to each and every original image in the training set to generate triple augmented images for every original input image. Thus, the number of training samples increased significantly, while the original stratification of the classes was maintained. We provide the list of properties employed for data augmentation in the following:

Table 3.1: Image Augmentation Parameters Used in the Study

Augmentation Type	Value/Range
Rotation Range	20 degrees
Width Shift Range	0.2 (20% of image width)
Height Shift Range	0.2 (20% of image height)
Shear Range	0.15
Zoom Range	0.2
Horizontal Flip	Enabled
Brightness Range	0.7 to 1.3
Fill Mode	Nearest Neighbor

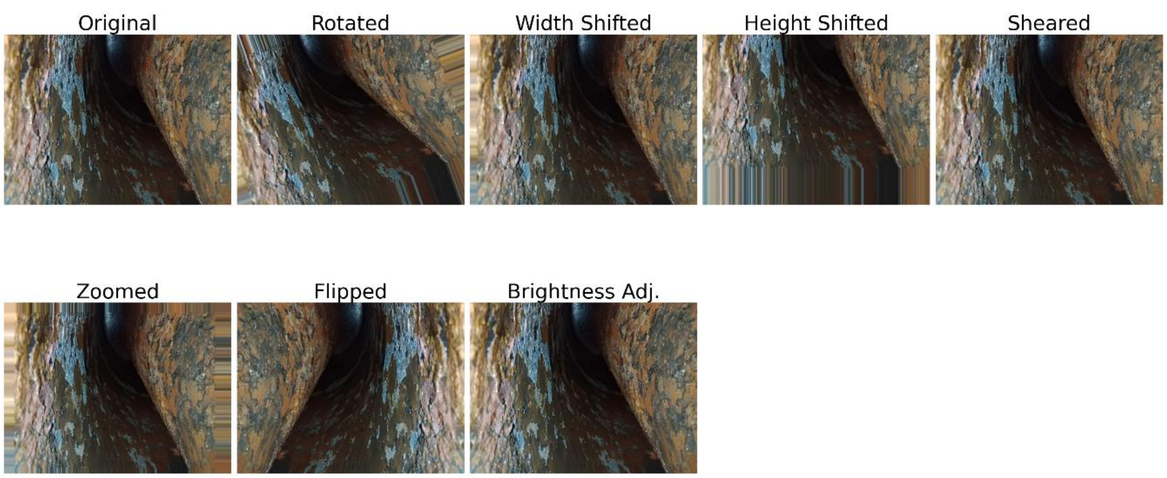


Fig. 3.3: Sample image and its augmented versions using the defined augmentation pipeline.

3.3.3 Image Preprocessing

In order to satisfactorily preprocess the acquired images for training with a deep learning model, all images were resized to 224 by 224 pixels. This resizing made the input shape uniform across all images in the dataset, to help convolutional neural networks such as Inception V3 and VGG16, which are typically used for industrial image classification tasks, work best.

A variety of preprocessing methods were tested in an effort to enhance image quality and separation. Each method was utilized and evaluated with training the Inception V3 network for 10 epochs and then testing its performance based on various metrics. The preprocessing methods compared were:

CLAHE: Contrast Limited Adaptive Histogram Equalization (CLAHE) enhances local contrast of the image, particularly in low dynamic range areas. The histogram equalization is applied to the localized grid tiles and after that these are combined.

Equation (for pixel p's histogram equalization):

$$p' = \frac{CDF(p) - CDF_{min}}{1 - CDF_{min}} \times (L-1) \tag{1}$$

$CDF(p)$in Equation 1 is the cumulative distribution function at pixel intensity p, L is the range of possible intensity levels.

Contrast Enhancement: In this method the intensity level is scaled to highlight edges and patterns and then image is shifted to 8bit scale (0–255).

$$p' = \left((p - p_{min})/(p_{max} - p_{min})\right) * 255 \tag{2}$$

where p' (Equation 2) is a pixel value, p_{min} and p_{max} are the minimum and maximum pixel intensity, respectively.

Gamma Correction: Gamma correction nonlinearly changes the brightness to emphasize features in the darker or lighter areas.

$$p' = 255 \times \left(\frac{p}{255}\right)^{\gamma} \tag{3}$$

where γ (Equation 3) is the gamma (commonly 1 to darken).

Noise Reduction: This approach employs denoising filters (e.g., median or Gaussian blur) after enhancement to reduce random noise, which may introduce in the real-world factory images.

$$G(x,y) = \frac{1}{2\pi\sigma^2}\, e^{-\frac{x^2+y^2}{2\sigma^2}} \tag{4}$$

In Equation 4, $G(x,y)$: The value of the Gaussian kernel at position (x,y), σ: Standard deviation (controls the "spread" of the blur), (x,y): Coordinates in the kernel

VGG16 Preprocessing: This applies the VGG16's default preprocessing function (subtracting the mean based on the ImageNet dataset):

$$p' = p - \mu \tag{5}$$

where μ (Equation 5) is the channel-wise mean (e.g., [103.939, 116.779, 123.68] for BGR).

The following preprocessing pipelines were tested:

i. Grayscale + CLAHE (Contrast Limited Adaptive Histogram Equalization)
ii. Contrast Enhancement + Unit 8 Conversion
iii. Unit 8 Conversion + Enhancement + Gamma Correction
iv. Unit 8 Conversion + Enhancement + Noise Reduction
v. VGG16 Preprocessing (using the built-in preprocessing function for VGG16)

Each pipeline was assessed based on training and testing accuracy, loss, and 10-fold cross-validation accuracy. The best-performing approach in terms of balanced generalization and training performance was Unit 8 Conversion + Enhancement, which offered a strong combination of accuracy and low loss.

3.3.4 Feature Extraction for Machine Learning (ML) Models

Different feature extraction techniques were attempted to produce a robust pipeline in the machine learning for boiler scale ML classification. We desired to uncover which method can provide the most discriminative description, and how to describe the scale patterns and the surface textures of the industrial images more effectively. Inspired by the literature review, we proposed complexities in feature extraction algorithm also held in combination with KNN and best executed feature extraction algorithm with trade-off among training, cross-validation, and test accuracy was selected. We include both, classical image processing and deep learning-based solutions in the spectrum of employed feature representations, which involves different image properties.

HOG (Histogram of Oriented Gradients): Captures the gradient structure and edge orientation, ideal for identifying shapes and patterns. Equation 6 and 7 for Gradient magnitude and orientation:

$$G_x = I(x+1, y) - I(x-1, y), \quad G_y = I(x, y+1) - I(x, y-1) \tag{6}$$

$$\text{Magnitude} = \sqrt{G_x^2 + G_y^2}, \quad \theta = \tan^{-1}\left(\frac{G_y}{G_x}\right) \tag{7}$$

GLCM (Gray-Level Co-occurrence Matrix): Analyzes texture by computing how often pixel pairs with specific values and spatial relationships occur. Key Features equations 8,9 which used to calculate Contrast, Homogeneity

$$\text{Contrast} = \sum_{i,j} |i-j|^2 P(i,j) \tag{8}$$

$$\text{Homogeneity} = \sum_{i,j} \frac{P(i,j)}{1+|i-j|} \tag{9}$$

LBP (Local Binary Pattern): Encodes texture by thresholding each pixel's neighborhood. And Equation 10 used for the LBP calculation.

$$LBP(x_c, y_c) = \sum_{p=0}^{P-1} s\left(i_p - i_c\right) \cdot 2^p \tag{10}$$

where x=1 if x≥0, else 0; i_c is the center pixel and i_p are the neighboring pixels.

VGG16 Feature Extraction: Uses intermediate layers of the pre-trained **VGG16** CNN to extract deep visual features. These representations encode semantic and abstract information useful for classification.

2DSCN (Two-Dimensional Supervised Contrastive Network): Learns feature representations by contrasting positive and negative sample pairs in a supervised manner. Contrastive Loss Function**:**

$$\text{L} = \sum(i,j) y_{(ij)} \cdot D_{(ij)}^2 + \left(1 - y_{(ij)}\right) \cdot max\left(0, m - D_{(ij)}\right)^2 \tag{11}$$

where $D_{(ij)}$ is the distance between feature vectors, $y_{(ij)}$ is 1 for similar pairs and 0 otherwise, and m is the margin.

SIFT (Scale-Invariant Feature Transform): SIFT extracts and describes scale-, rotation- and partly affine-invariant local features from images. It involves four key steps:

Scale-Space Extrema Detection: Candidate interest points are found by convolving the image with Gaussian filters at different scales and then computing the Difference of Gaussians (DoG),

$$D(x, y, \sigma) = L(x, y, k\sigma) - L(x, y, \sigma) \tag{12}$$

where $L(x, y, \sigma)$ is the image convolved with a Gaussian of scale σ, and k is a constant multiplicative factor.

Key point Localization: Key points are then detected by finding maxima/minima over DoG images, and refined based on Taylor expansion around the extremum:

$$\hat{x} = -\frac{\partial^2 D^{-1}}{\partial x^2} \cdot \frac{\partial D}{\partial x} \tag{13}$$

Orientation Assignment: An orientation is assigned to each key point in reference to the gradient direction of the local image patch:

$$\theta(x, y) = \tan^{-1}\left(\frac{L(x,y+1)-L(x,y-1)}{L(x+1,y)-L(x-1,y)}\right) \tag{14}$$

2DSCN + SIFT Combination: Combines the strengths of both deep contrastive learning and local feature descriptors, yielding rich multi-scale representations.

3.4 Machine Learning Models for the Image Classification

After feature extraction, several traditional machine learning classifiers were trained and tested for the optimal classifier in boiler scale image classification. The chosen models are representative of an extensive range of learning strategies, due to their ensemble-based and classical learning methods.

The models under consideration are the:

- XGBoost
- Logistic Regression
- LightGBM
- K-Nearest Neighbors (KNN)
- Support Vector Machine (SVM)
- Random Forest
- Gradient Boosting
- AdaBoost
- Decision Tree
- Naïve Bayes

The optimal classifiers were evaluated on features derived from the top performing pipeline and compared to select the best performing classifier for deployment.

3.4.1. Hyperparameter Tuning ML Model

To at least maximize the top performing machine learning model, a full hyperparameter tuning process was carry out through Randomized Cross-Validation (Randomized CV). This method has the capacity to efficiently search through a high-dimensional parameter space (which is expensive to evaluate), without relying on naive grid search.

Random 900 combinations were picked and tested. Randomization of the approach allowed for parameter space to be further explored, therefore increasing the likelihood of the discovery of a globally optimal parameter set.

The obtained best parameters as suggested Based on guideline Extracted from this experiment were further used to fine-tune our model on the training and secret sets and assess its prediction capabilities on a test set.

3.5 Deep Learning Model Construction

In order to enhance classification performance in addition to on-the-shelf machine-learning methods, various transformer-based and convolutional deep learning architectures were explored. For each model, performance and computational effectiveness were considered. The implemented models are as follows:

EfficientNet V2L:

EfficientNet V2L represents a high-end model in the EfficientNetV2 family. It build on top of the previous works by adding Fused-MBConv layers for more expedited execution and better parameter efficiency. It also benefits from progressive learning strategy where we feed training input images of increasing resolution. This architecture is suitable for large-scale image classification, with efficient training time and enhanced accuracy.

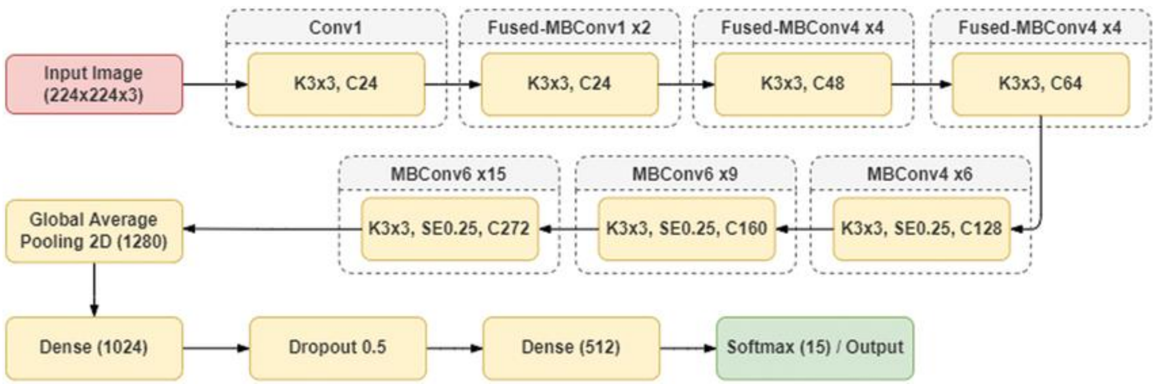


Fig. 3.4: EfficientNet V2L architecture

MobileNet V2:

MobileNet V2 has been designed with a focus on edge devices and mobile. It uses inverted residual blocks and linear bottlenecks to allow the model to keep high performance, while also reducing the size and latency. It has a lightweight structure, making it suitable for realtime applications on resource constrained devices.

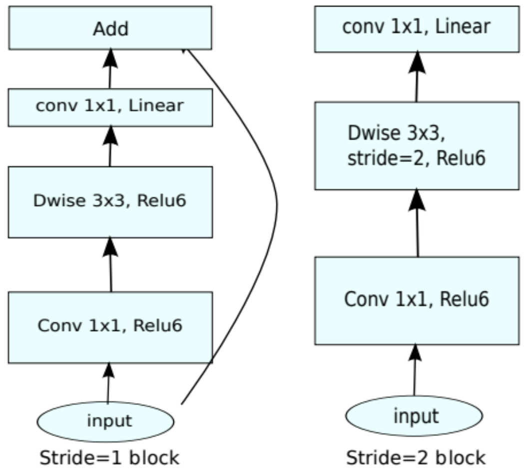


Fig. 3.5: MobileNet V2 architecture

EfficientNet B0:

EfficientNet B0 is the baseline model of the EfficientNet family. It systematically scales the network dimensions such as depth, width, and resolution in a compound manner. By using MBConv blocks and swish activation it achieves good performance with less number of parameters than classic models.

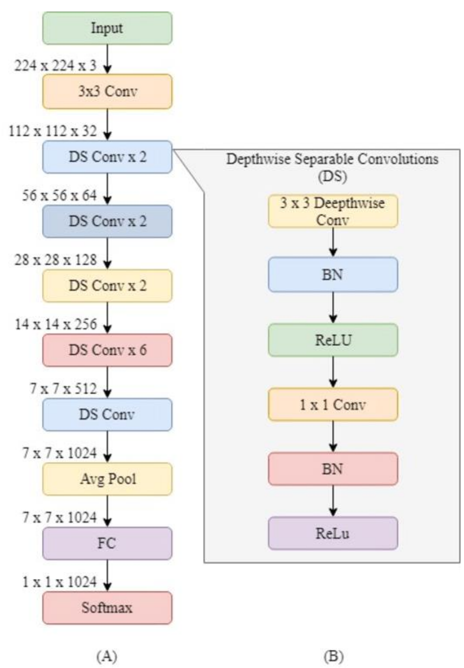


Fig. 3.6: MobileNet B0 architecture

VGG16:

VGG16 is simple and deep. It places multiple 3x3 CONV layers in a uniform block architecture with then FC layers. Although it is relatively heavy computationally, it presents a natural and understandable structure being widely used as a benchmark in classification problems.

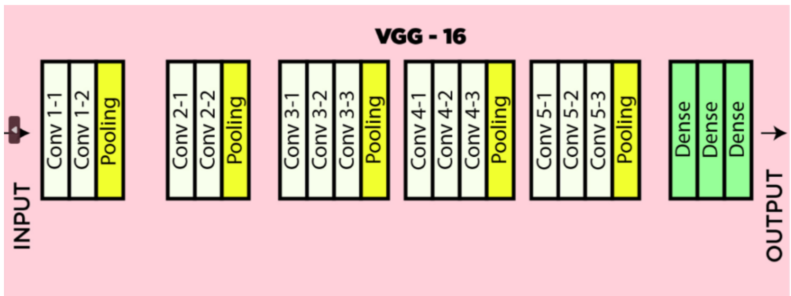


Fig. 3.7: VGG16 architecture

Xception:

It can be regarded as an extension of the Inception family, in which Inception-type of modules would be replaced with depth wise separable convolutions, without internes. The resulting model preserves high accuracy with fewer computations and fewer parameters so that it can be scaled up efficiently to larger scale data.

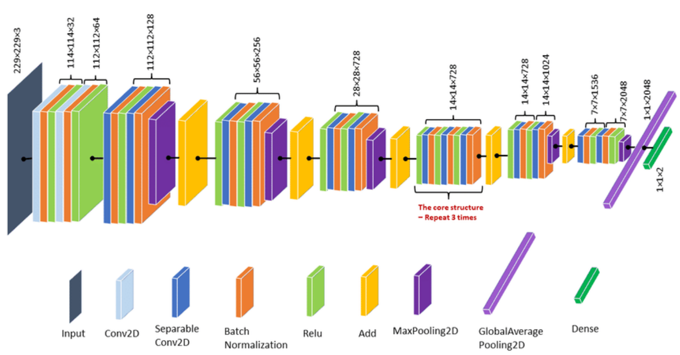


Fig. 3.8: Xception architecture

ResNet-18:

ResNet-18 develops the concept of residual connection, which facilitates stable training of deeper networks, in the sense that it helps to alleviate the vanishing gradient issue. Every residual block enables the model to "skip" connections which allows information to pass through to various different layers. This allows intermediate representation learning without loss, even with relatively deep networks.

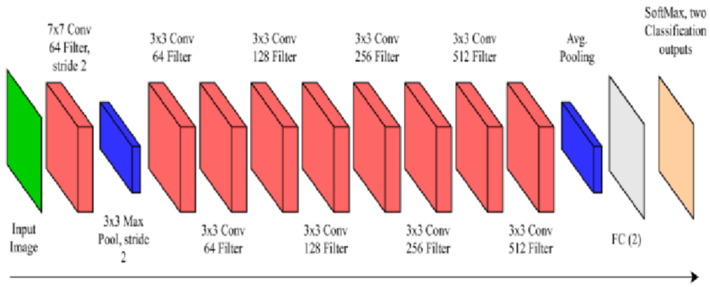


Fig. 3.9: ResNet 18 architecture

3.6 Customized Model Development

In order to explore the state-of-the-art performance and alleviate the drawbacks of existing off-the-shelf architectures, we construct a novel deep model by combining multiple advanced modules, including CNN, Attention Mechanism, Squeeze-and-Excitation (SE) Network and Residual Fusion Strategy (RFS). This hybrid architecture was designed to capture global and local discriminative features from boiler scale images by combing the advantages of deep residual learning and dynamic feature recalibration.

3.6.1 Model Components

Convolutional Layers (CNN Backbone): The architecture of the model is built on convolutional layers to model spatial hierarchies in the image. At each layer, convolutional filters are used to obtain local patterns:

$$F(i,j) = (I * K)(i,j) = \sum_m \sum_n I(i+m, j+n) \cdot K(m,n) \tag{15}$$

Where, I is the input image and K is the convolutional kernel.

Squeeze-and-Excitation (SE): The Squeeze-and-Excitation (SE) blocks follow the architecture of ResNet and are closely followed by other state-of-the-art methods to improve representation in CNN's and have been proven effective. SE blocks re-weight feature maps by capturing inter-channel dependencies. The Global Average Pooling is performed for the "squeeze" operation as follows:

$$z_c = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} X_c(i,j) \tag{16}$$

with two fully connected layers for learning the excitation weights sc that are employed to re-scale the input features:

$$\widetilde{X_c} = s_c \cdot X_c \quad (17)$$

Residual Fusion Strategy (RFS): The proposed architecture considers the features from Fusion unit and Attention Unit. Residual connections ease the task of training deep models, due to the improved flow of gradients and the possibility to reuse feature information. Each residual block is formalized as:

$$y=\mathrm{ReLU}(F(x)+x) \quad (18)$$

where F(x) is the transformation composed of convolution, batch normalization, and SE block layers.

Custom Attention Mechanism: A trainable attention layer is added to highlight the most informative spatial or channel-wise regions. The saliency scores are calculated as:

$$A=\mathrm{softmax}(XW) \quad (19)$$

where rows of A are normalized into a probability distribution.

where X indicates the feature input, and W denotes a learned attention weight matrix. The output is:

$$X'=A\cdot X \quad (20)$$

Fully Connected Layers: After deep feature extraction and attention refinement, the representation is flattened, and fed into dense layers followed by softmax activation for multiclass classification:

$$F(x) = \mathrm{softmax}(Wx+b) \quad (21)$$

3.6.2 Hyperparameter Optimization

The optimization of architecture configuration was carried out based on automated hyperparameter search employing Keras Tuner. We sampled a total of 80 trials using Hyperband search strategy, searching for different combinations of:

- Number of Convolution filters ([32, 64, 128])
- Dense blocks (([64, 96, 128, 160, 192, 224, 256]))
- Dropout rates (0.2 to 0.5)
- E_learning rates(1e-4, 1e-3, 1e-2)

Each trial was trained until 50 epochs with early stopping to avoid overfitting. The total tuning time spent took around 9 hours to compute. The model configuration with best result was based on the validation accuracy.

3.7 Ablation Study

In order to test the performance of competitive learning models, traditional ML classifiers, pre-trained DL models and the proposed fine-tuned models, a thorough analysis was done to check which was the best competitive model after everything. For this, the performance of the model with various indexes were compared, in order to find the most simple and precise classification methods employed in the industrial scale classification of the boiler.

The criteria and methods of evaluations used were:

3.7.1 Accuracy of training and testing: Accuracy determines what proportion of the instances are accurately predicted as follows:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \tag{22}$$

Where, TP: True Positives, TN: True Negatives, FP: False Positives, FN: False Negatives. The training and testing accuracies were compared to find overfitting or underfitting.

3.7.2 Training and Testing Loss: The prediction error is measured by the so-called loss functions. For multi-class classification, the Categorical Cross-Entropy Loss was employed:

$$L = -\sum_{i=1}^{N} \sum_{c=1}^{C} y_{i,c} \cdot \log\left(\widehat{y_{i,c}}\right) \tag{23}$$

Where $y_{i,c}$ is the true label and $\widehat{y_{i,c}}$ is the predicted probability of class c for sample i.

3.7.3 Prediction Time: We measured the inference speed of all the models, i.e., the total time of predicting all the test samples. Reduced prediction time drives the algorithm in the direction to hard-real-time or near-real-time applications for the industrial district.

3.7.4 Per-class Accuracy - Individual Class: To analyze if a model was better able to work on individual classes (e.g, Calcium Carbonate, Iron Oxide etc.), the model's per-class accuracy was taken to see if there was a class imbalance or model bias between different scale types.

3.7.5 Cross-validation accuracy- Average value: The stability and transferability of the prediction model were assessed using 10-fold cross-validation. Error-rate averaged over all folds.

3.7.6 Confusion Matrix: Confusion matrix is a table that is referenced to analyze the performance of a classification algorithm (or classifier) wherein each row of the matrix represents the instances in a predicted class while each column represents the instances in an actual class. It is interesting to reveal some trend of misclassification.

3.7.7 AUC-ROC Curve (Area Under Receiver Operating Characteristic): For multi-class evaluation, macro-averaged AUC-ROC scores were computed. The ROC curve plots True Positive Rate (TPR) vs False Positive Rate (FPR) at various thresholds:

$$\mathrm{TPR} = \frac{TP}{TP+FN}, \quad \mathrm{FPR} = \frac{FP}{FP+TN} \tag{24}$$

The AUC (Area Under the Curve) score quantifies the classifier's ability to distinguish between classes.

3.7.8 Training Curves: For each deep learning model, we plotted the training and validation accuracy/loss curves over epochs to investigate if the deep learning model was convergent and verify concerns such as over-fitting, under-fitting, etc. This multi-metric evaluation scheme offered a global view about sacrifices on efficiency-competence-accuracy-speed trade-offs, that helped to select the rightest own model under deploy circumstances.

3.8 Explainable AI Using LIME

For explanations of the predictions of both traditional machine learning and deep learning models we used LIME (Local Interpretable Model-agnostic Explanations) as a post-hoc explainability method. This was important to provide some explanation of how individual predictions were reached, which is particularly crucial in industrial setting, since interpreting why a scale is predicted to have a certain class is important to decide the maintenance action.

Subsequently, LIME estimates a complicated non-linear model around any prediction by a local model (often linear or decision tree). This local surrogate model aids understand what regions of the input (i.e., image regions or pixel intensities in our case) were most responsible for the final prediction.

CHAPTER 4

EXPERIMENTAL RESULTS AND DISCUSSION

4.1 Preprocessing Technique Selection

Comparison of different pre-processing steps used prior to training of the models is presented in Table 1 using training accuracy, training loss, 10-fold cross-validation accuracy, test accuracy and test loss. Among the five techniques tested, Contrast Enhancement and Unit 8 conversion was the most balanced and flexible approach. It achieved the high training accuracy 0.9688 and the low training loss 0.0684, indicating the nice learning process. Most importantly, the method reached a higher level of performance

Table 4.1: Performance Comparison of Different Preprocessing Techniques

Preprocessing Technique	**Train Accuracy**	**Train Loss**	**10-Fold Cross-Validation Accuracy**	**Test Accuracy**	**Test Loss**
Grayscale + CLAHE	0.937	0.1402	0.9062	0.8485	0.4507
Contrast Enhancement + Unit 8 Conversion	0.9688	0.0684	0.9306	0.8687	0.3964
Unit 8 + Enhancement + Gamma Correction	0.9375	0.1160	0.9097	0.8485	0.4229
Unit 8 + Enhancement + Noise Reduction	1.000	0.0227	0.9167	0.8586	0.4760
VGG16 Preprocessing	0.9688	0.0344	0.9141	0.8384	0.5935

than others in term (best of both AUC ROC (0.711) and 10 cross cross-validation (0.9121), best overall testing ROC (0.8687), and lowest overall testing loss (0.3964), which depicted the privileged generalization performance. In contrast, the Noise Reduction approach fitted the training set perfectly (1.000), but it unsatisfactorily generalized to the test, indicating an overfit. Estimates were lower and test loss was higher for the other methods (Grayscale + CLAHE, Gamma correction and VGG16 Preprocessing).

Therefore, Contrast Enhancement + Unit 8 Conversion was selected as pre-processing method for all the remaining model training and testing.

4.2 Feature Extraction Method Selection for ML Based Prediction

Feature extraction was conducted after implementation of preprocessing method using popular K-Nearest Neighbors (KNN) model. The results of different feature extraction methods are shown in Figure 4.1.

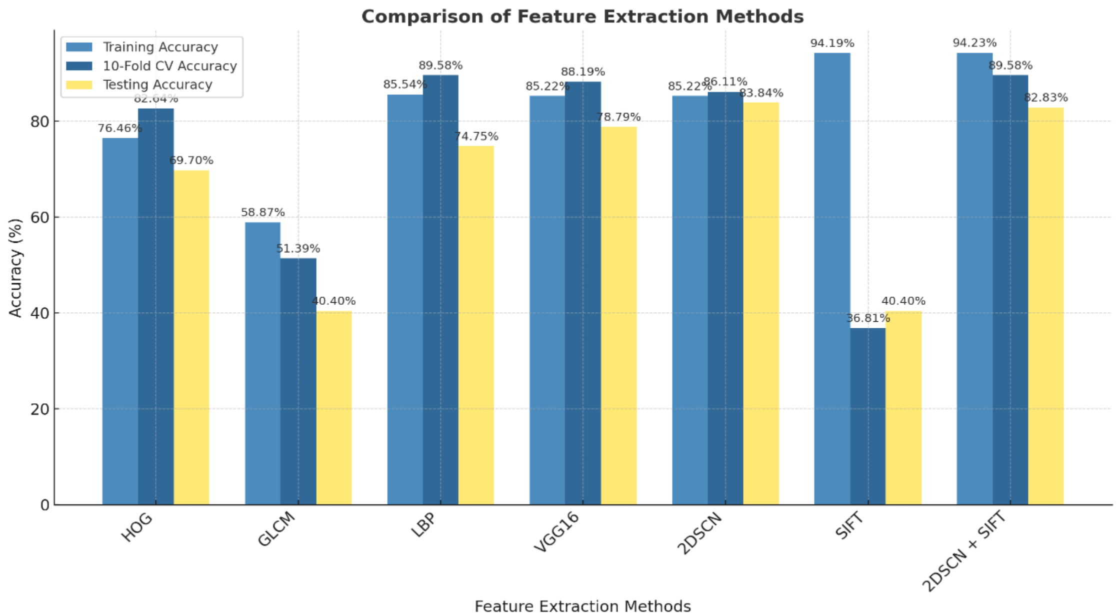


Fig 4.1: Feature Extraction Methods Results

Figure 4.1 compares the job of seven different feature extraction methods HOG, GLCM, LBP, VGG16, 2DSCN, SIFT and a combination of 2DSCN and SIFT with regards to Training Accuracy, 10-Fold CrossValidation Accuracy, Testing Accuracy. This comparison reveals the consistency and generalization ability from one evaluation stage to another of both methods.

Especially for 2DSCN + SIFT, it obtains the best results of 94.23% training accuracy, 89.58% cross-validation accuracy and 82.83% testing accuracy, indicating that it has strong learning capacity as well as generalization to un-seen ones. In the same vein, SIFT on its own resulted into the best training accuracy (94.19%) but fared poorly at cross validation (36.81%) and testing (40.40%) accuracy and possibly overfitting.

Baseline approaches such as HOG and LBP had intermediate performances having 74.75% testing accuracy for LBP, 69.70% for HOG and presented also better generalization than GLCM. GLCM, achieved inferior performance in all three cases, also evidencing restricted usefulness in this type of analysis.

Finally, deep learning- based features extracted from VGG16 and 2DSCN demonstrated good balance, with VGG16 exhibiting, respectively, 88.19% and 78.79% cross- validation and testing accuracies, hence they may serve as strong standalone methods.

Overall, the hybrid model (2DSCN+SIFT) achieved best performance compared with the rest methods, indicating that combining the advantages of hand-crafted feature and learned feature can be suitable for complex feature representation in ICA.

4.3 Machine Learning Models Performance Comparisons

In this section, we thoroughly evaluate the models we built for boiler scale types classification. Performance evaluation is made concerning required metrics such as the overall classification accuracy, class-level behavior, prediction effectiveness, and the diagnostic visualization. These results reveal differential predictive and utility tradeoffs of models.

4.3.1 Overall Model Performance

A brief comparison of the results for the eleven machine learning models is shown in (Table 4.2) framed for the training and testing values of the accuracy, precision, recall, and F1 score. The results are as follows: Tuned XGBoost model achieved the highest classification accuracy of 90.12% and with an F1 of 0.9009. All these observations imply a relatively good generalization property and less overfitting since the training and testing accuracies are so near. Other models XGBoost, Logistic Regression and LightGBM (in decreasing order) were also competitive but with a lower test accuracy value of 88%-89%. In contrast, Naive Bayes and Decision Tree have shown poor generalization ability with test accuracies below 80% implying that they may struggle to encapsulate overlapped or complex patterns features.

Table 4.2: Performance Comparison of Different ML models

Model	Train Accuracy	Test Accuracy	Precision	Recall	F1 Score
Tuned XGBoost	0.9993	0.9012	0.9019	0.9012	0.9009
XGBoost	0.9993	0.8889	0.8889	0.8889	0.8882
Logistic Regression	0.9037	0.8848	0.8844	0.8848	0.8844
LightGBM	0.9993	0.8848	0.8858	0.8848	0.8841
K-Nearest Neighbors	0.9423	0.8683	0.8678	0.8683	0.8679
SVM	0.88	0.8642	0.8693	0.8642	0.8638
Random Forest	0.9993	0.8601	0.8601	0.8601	0.8592
Gradient Boosting	0.9774	0.8601	0.8611	0.8601	0.8592
AdaBoost	0.8554	0.8148	0.8212	0.8148	0.8127
Decision Tree	0.9993	0.7942	0.7952	0.7942	0.7946
Naive Bayes	0.757	0.7407	0.751	0.7407	0.736

These results demonstrate how the better hyperparameter tuning and ensemble learning can improve model robustness and performance.

4.3.2 Reliability of Cross-Validation

Model stability was estimated with cross-validation analysis on various data splits. The mean classification accuracies are displayed in Figure 4.2. Tuned XGBoost was still the best model with 92.91% as the average accuracy in cross-validation, which indicates its robustness to the subset of data. A final note is that the other models (XGBoost and Logistic Regression) did not even change much at all, goes to show how strong those models are. The performance agreement between CV and test data also presents these models as ready for real-world roll-out.

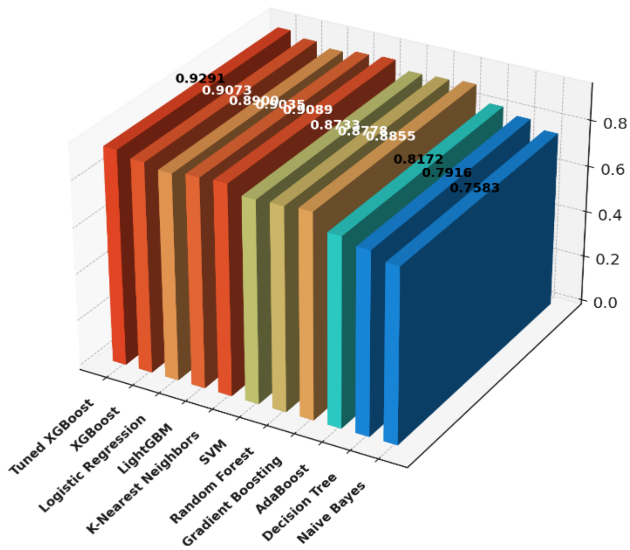


Fig 4.2: 10-Fold Cross-Validation Accuracy of ML Models

Such a high level of constant prediction has much value in applications where it is important to have reliable predictions.

4.3.3 Trends in Class-Wise Accuracy

Figure 4.3 presents the class-wise accuracies of the three types for Calcium Carbonate, Iron Oxide, and Others, to yield clear interpretable information on the model behaviour. Analogous tuning of XGBoost however, led to balanced and higher performance across all the three classes (90%, 95% and 85% accuracy respectively). Crucially, many of the (particularly Naive Bayes) models seem to do very badly on the Others class, hinting at a probable high degree of class imbalance and/or feature overlap. This class-based breakdown is important to interpret in real world scenarios, particularly when misclassification of small classes (e.g., Others) can have operational consequences.

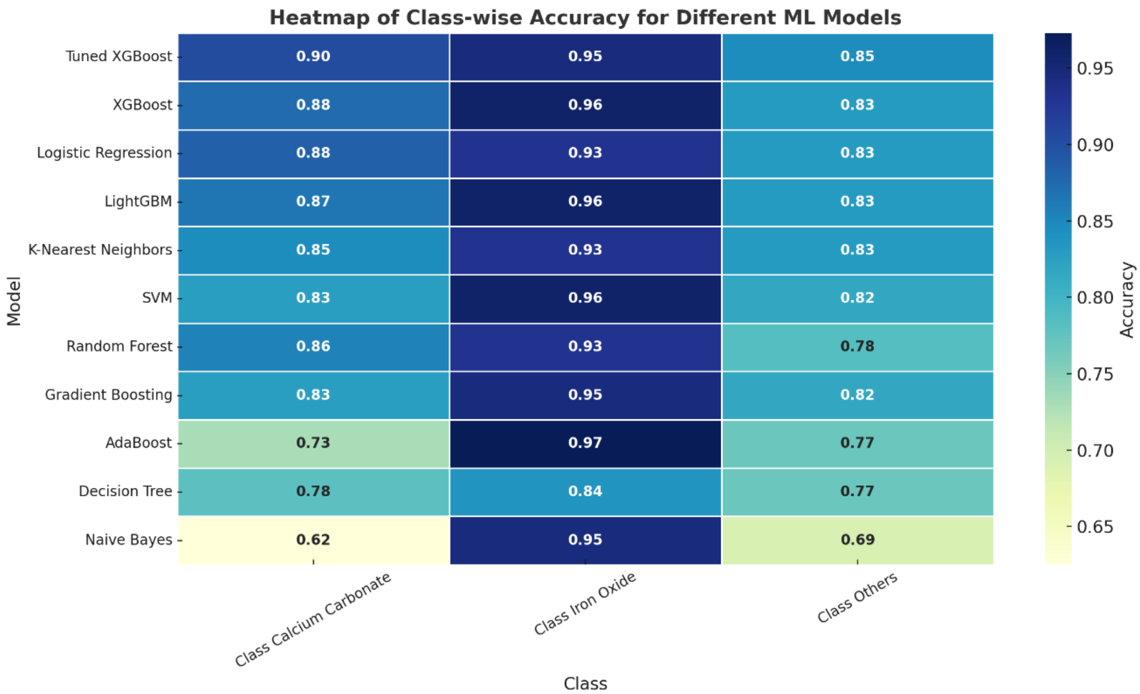


Fig 4.3: Class-wise Accuracy in Different ML Models

4.3.4 Prediction Time Comparison

In addition, computational efficiency needs to be taken into account other than accuracy, especially for real-time scenarios. To visualise how the models from Figure 2 perform with respect to prediction time, in Figure 3, we plot avg. prediction time per test instance. As expected, simpler models like Logistic Regression, Decision Tree and Naive Bayes achieved the fastest predictions – < 1.5ms Tuned XGBoost on the other hand was relatively slow (in the order of 16 ms), which however is still very cheap given the superior predictive power. In comparison, SVM and AdaBoost provided the worst response time which might limit their use on real-time usage scenarios.

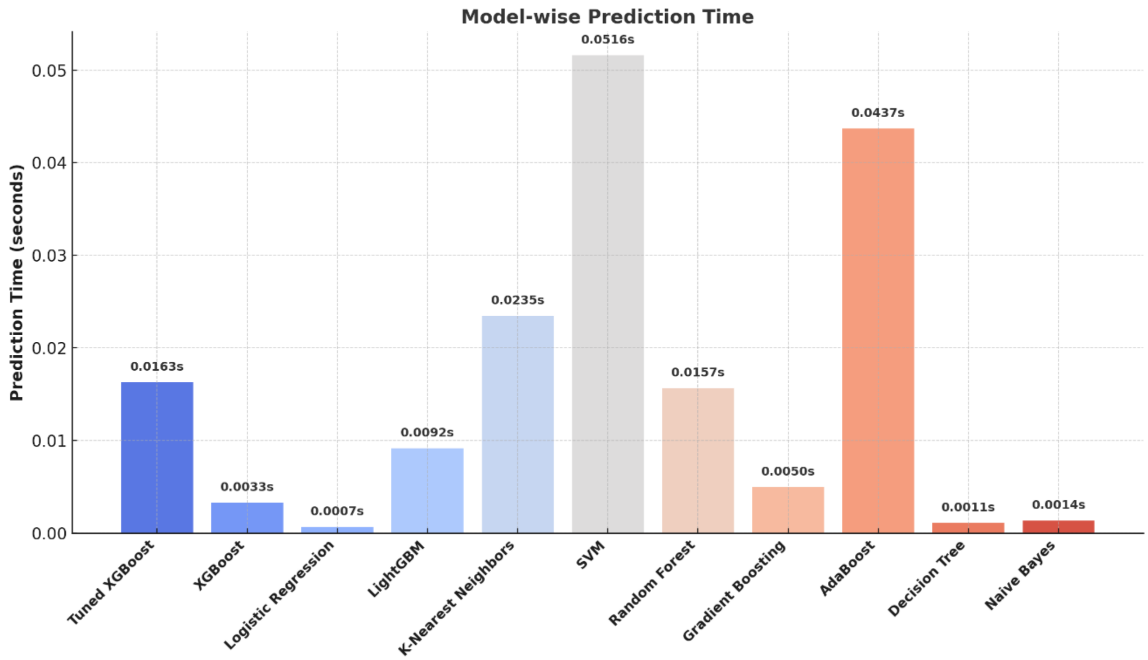


Fig 4.4 Model-wise Prediction Time on Test Data

These results will assist in applications where speed is important along with accuracy.

4.3.5 Confusion Matrix Interpretation

To get a closer look of the classification results confusion matrices of 4 best models are shown in Figure 4.5. These plots make it clear for how many times each machine has right/false predicted one class. The best separation of Class was observed for Tuned XGBoost with minimal misclassification for Calcium Carbonate and Others. Meanwhile, Logistic Regression and LightGBM misclassified Calcium Carbonate to Others a little more, which is possible the more likely number of high basis-precision failures in significantly imbalanced cases.

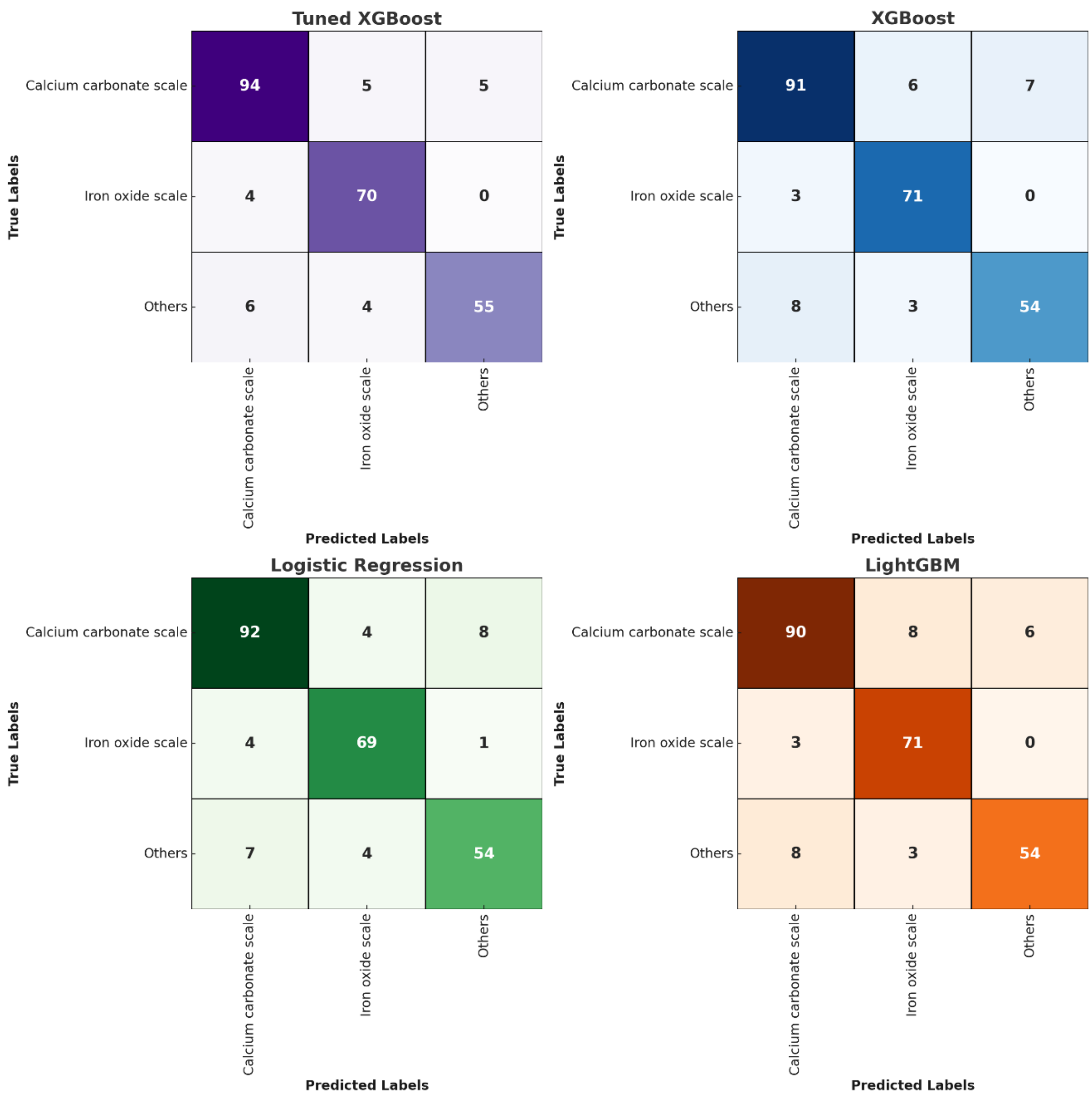


Fig 4.5: Confusion Matrix of Top 4 ML Models

This fine-grained perspective is crucial to understanding the real-world risks of false positive and false negative responses.

4.3.6 ROC-AUC Curve Evaluation

Lastly, AUC values and ROC curves determine model discrimination (Fig. 4.6). The optimized XGBoost showed excellent class separability with results near to the AUC optimal (0.97–0.98) in all classes. XGBoost and LightGBM were close to very good while Logistic Regression slightly lower but still very strong AUC results, especially for Others.

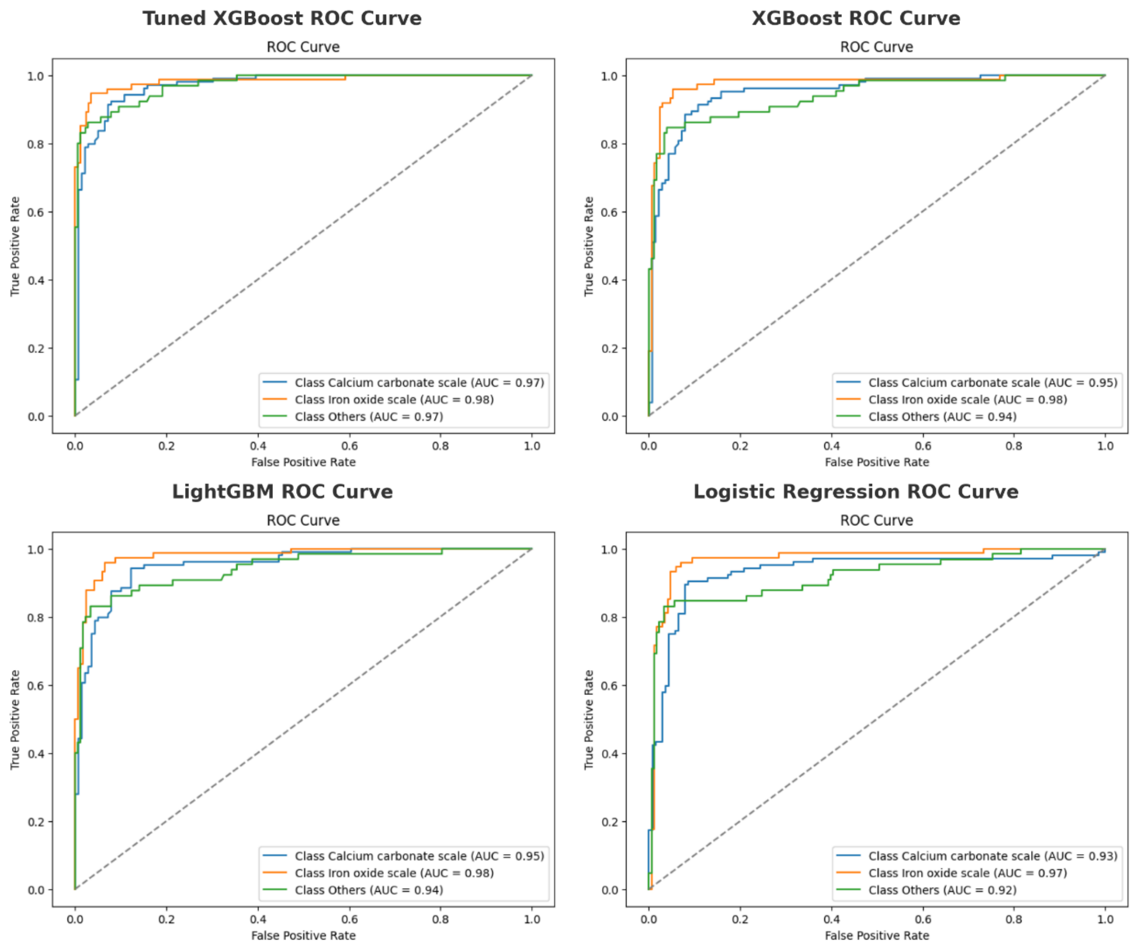


Fig 4.6: ROC curves of the leading 4ML models.

Such high AUCs indicate that the correct models predict, and they do so confidently.

Across all evaluation perspectives aggregate metrics, class-wise behavior, efficiency, and ROC confidence the Tuned XGBoost model consistently ranked highest. It combined predictive power with reasonable inference time and balanced class performance, making it the most reliable model for boiler scale classification tasks.

4.4 Deep Learning Models Performance Comparisons

This section provides a comprehensive experimentation on different deep learning networks for the industrial boiler scale classification problem. The goal was to discover the best model in terms of performance, stability and computational cost, based on various performance measures including classification accuracy, learning dynamics, cross-validation consistency, class-wise behavior, prediction speed and model interpretability. We provide detailed analysis for best model (ScaleNet V1 which is a custom-designed CNN) as well as its layer level composition and explainability insights.

Table 4.5: Performance Comparison of Different ML models

Model	Train Accuracy	Test Accuracy	Train Loss	Test Loss
ScaleNet V1	0.993	0.9342	0.0178	0.4036
EfficientNet V2L	0.9861	0.9012	0.0332	0.3907
MobileNet V2	0.9774	0.8971	0.0572	0.5345
EfficientNet B0	0.96	0.893	0.2439	0.4398
VGG 16	0.904	0.8848	1.1812	1.2533
Xception	0.8585	0.8765	0.3398	0.4593
Resnet 18	0.9604	0.8189	0.111	0.7094
Inception V3	0.5692	0.6255	0.9289	0.8225

4.4.1 Overall Model Performance

The performance of deep learning models in the training and test datasets are summarized in Table 4.3. Of the eight models evaluated, including both standard architectures and this paper's custom ScaleNet V1 model, performance is evaluated by training and test accuracy, as well as the associated loss values.

ScaleNet V1 was the best performing model, with a testing accuracy of 93.42% and an extremely low training loss (0.0178). Furthermore, the ROC curve for all classes ranges from 0.9827 to 0.9950 (Figure 4.7), which means good generalization is achieved during training. Followed by EfficientNet V2L and MobileNet V2 who only narrowly trailed with test accuracies of 90.12% and 89.71% respectively. On the other hand, models like

ResNet18 and Inception V3 still showed quite low test accuracy, particularly in dealing with diversification of inputs.

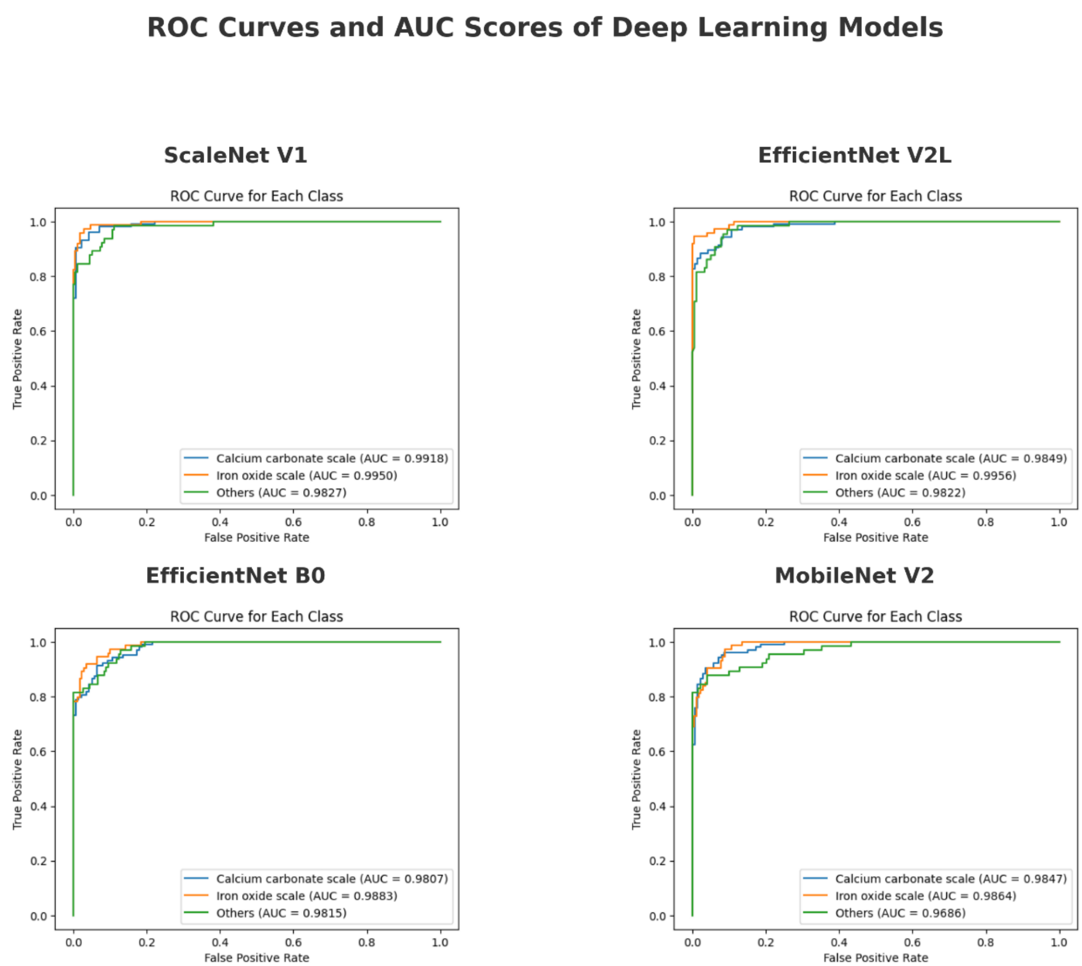


Fig 4.7: ROC Curve Top Four Deep Learning Models

4.4.2 Training Accuracy and Loss Curves

The evolution of the training accuracy and loss over epochs reveals the learning behaviors and stability of the models. Training curves of the best 4 models are shown in Figure 4.8. Both ScaleNet V1 and EfficientNet V2L showed stable convergence and almost completely optimal accuracy and only small loss fluctuations. MobileNet V2 converged well and EfficienNet B0 appeared less stable initially but later became stable.

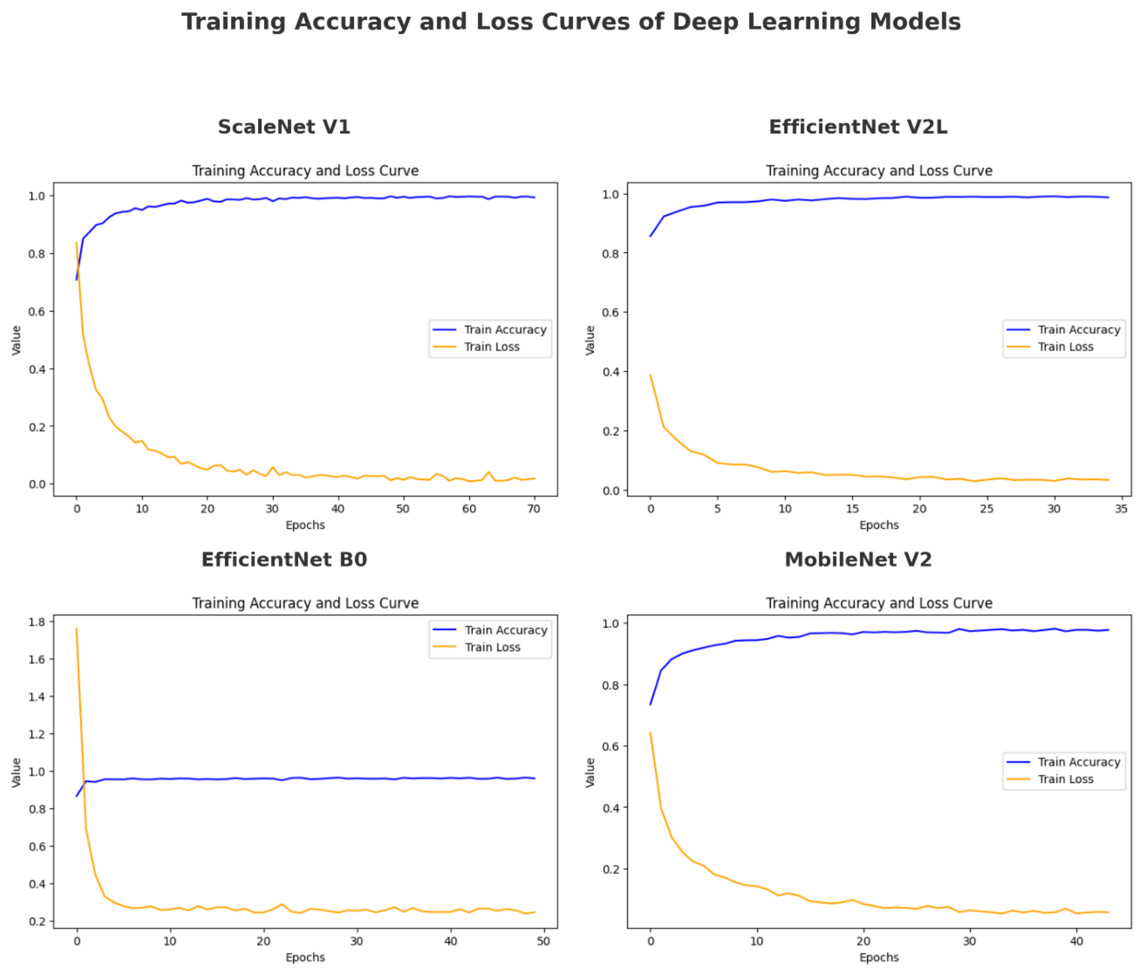


Fig 4.8: Training Accuracy and Loss Curves of Top Four Deep Learning Models

These curves validate the training efficacy of the models and highlight the early learning patterns that contributed to their final performance.

4.4.3 Cross-Validation Accuracy

Finally, 10-fold cross-validation was employed to verify the robustness of the model on various data distributions. Average cross-validation accuracy for each model is presented in Fig 4.9. The highest mean accuracy was obtained by MobileNet V2 (99.48)%, followed by ScaleNet V1 and EfficientNet V2L (both 98.89)%. These findings confirm the robustness and reproducibility of the best models.

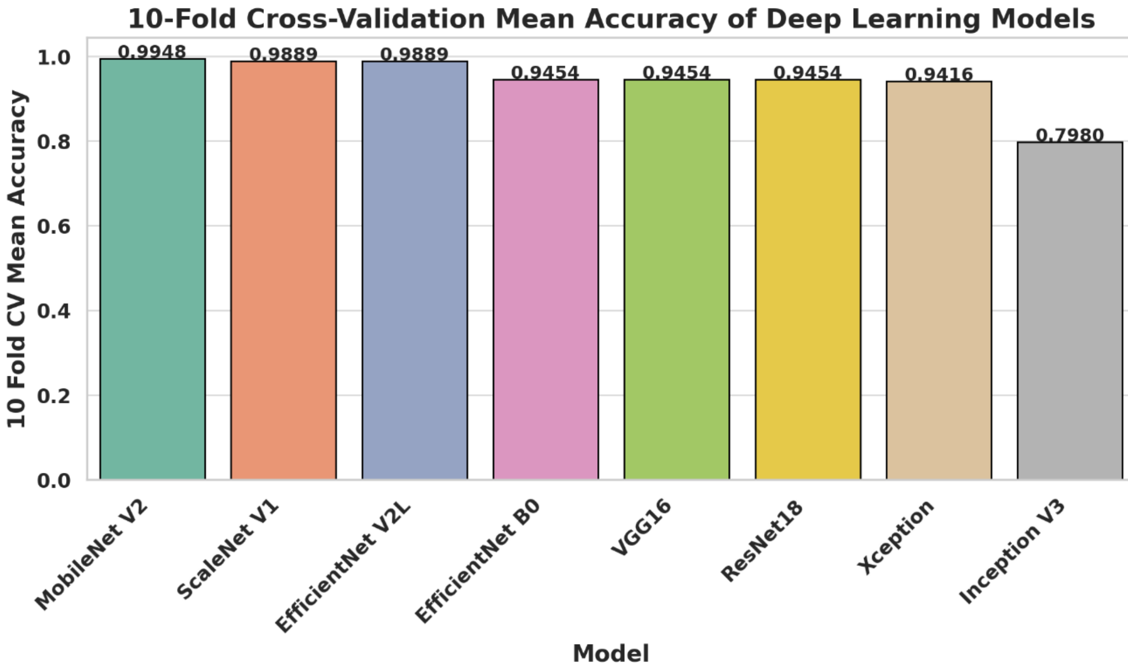


Fig 4.9: 10-Fold Cross-Validation Mean Accuracy of Deep Learning Models

4.4.4 Class-wise Accuracy Analysis

We include a class-wise accuracy heatmap over the three scale type categories in Figure 4.10. ScaleNet V1 demonstrated the most balanced and the highest accuracy in every class, i.e., 98.08%, 94.59%, and 84.62%. This further validates the effectiveness of the model in learning various patterns.

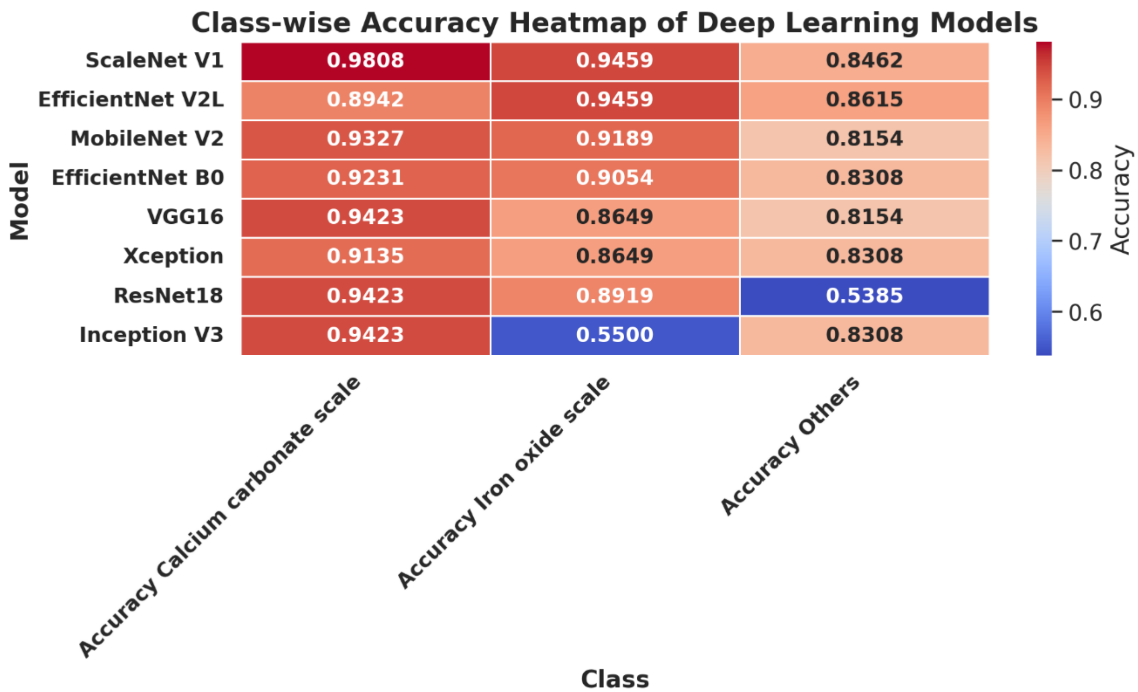


Fig 4.10: Class-wise Accuracy Heatmap of Deep Learning Models

Other models, such as ResNet18 and Inception V3, experienced significant drops in accuracy for the "Others" category, highlighting their limitations in handling less dominant classes.

4.4.5 Prediction Time Comparison

The speed of model inference is very important when it comes to the deployment in the real-life industrial systems. Figure 4.11 shows the inference time of each model on our test data. The mobile nets EfficientNet B0 and ScaleNet V1 were the fastest with 2.01 and 2.02 seconds, respectively. On the other hand the ResNet18 was the slowest at 11.16 seconds and may not be the most appropriate for time-critical applications.

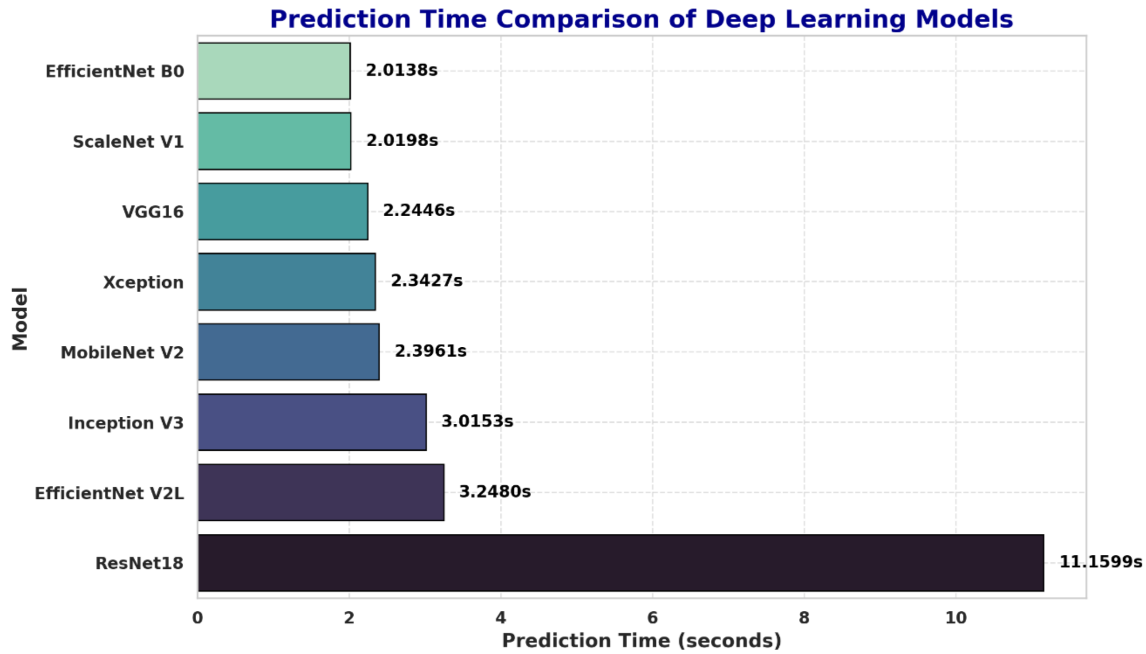


Fig 4.11: Prediction Time Comparison of Deep Learning Models

4.4.7 Confusion Matrix Evaluation

Confusion matrices were created to represent the distribution of correct and incorrect predictions over all classes. The matrices for the top four models are presented in Figure 4.12. Once again, ScaleNet V1 performed best with fewest misclassifications. For example, the number of misclassified samples in both the Calcium carbonate groups was 2 and that of "Others" was 3. This accuracy even strengthens the validity of our model.

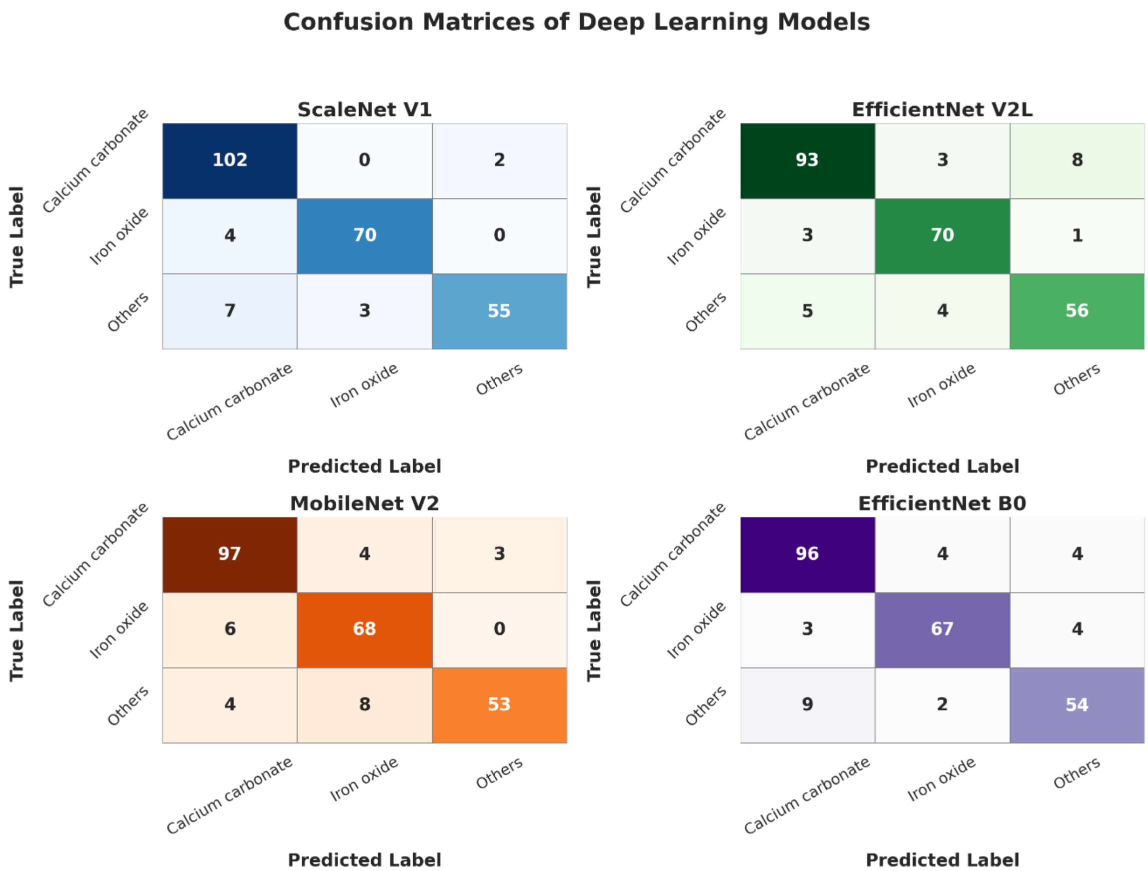


Fig 4.12: Confusion Matrices of Top 4 Deep Learning Models

4.4.8 Architecture of the Customized CNN Model (ScaleNet V1)

The most effective model, ScaleNet V1(Table 4 as layer structure), was developed using Keras Tuner for neural architecture search and hyperparameter optimization. This model incorporates multiple advanced components:

- Residual Blocks to facilitate deep feature propagation without vanishing gradients.
- Custom Attention Layer to enhance spatial focus on scale patterns.
- Batch Normalization and ReLU layers for improved convergence and non-linearity.
- GlobalAveragePooling and Dropout for regularization and dimensionality reduction.

Table 4.4: Layer Propertise of ScaleNet V1

Layer (Type)	Output Shape	Parameters
InputLayer	(None, 224, 224, 3)	0
Conv2D	(None, 224, 224, 64)	1,728
BatchNormalization	(None, 224, 224, 64)	256
ReLU	(None, 224, 224, 64)	0
MaxPooling2D	(None, 112, 112, 64)	0
Conv2D	(None, 112, 112, 128)	73,728
BatchNormalization	(None, 112, 112, 128)	512
ReLU	(None, 112, 112, 128)	0
MaxPooling2D	(None, 56, 56, 128)	0
ResidualBlock	(None, 56, 56, 128)	298,120
ResidualBlock	(None, 28, 28, 256)	929,040
ResidualBlock	(None, 14, 14, 512)	3,709,472
CustomAttentionLayer	(None, 14, 14, 512)	262,144
GlobalAveragePooling2D	(None, 512)	0
Dense	(None, 64)	32,832
Dropout	(None, 64)	0
Dense	(None, 3)	195

The model's total parameter count is optimized to ensure balance between learning capacity and computational efficiency.

4.4.9 Explainable AI (LIME)

In order to make the decisions of the model interpretable and as transparent as possible, LIME (Local Interpretable Model-agnostic Explanations) was used. LIME explanations for several test samples are given in Figure 4.13. These illustrative images emphasize the regions in the image that are most relevant to the classification decision. The yellow zones represent the pixels that most contribute to the prediction, while the black or white areas have made little contribution.

LIME validated that ScaleNet V1 captures the essential structural patterns like rust lines, surface textures, and deposits consistent with expert visual observations. This leads to increased user trust and facilitates model deployment within safety-critical applications.

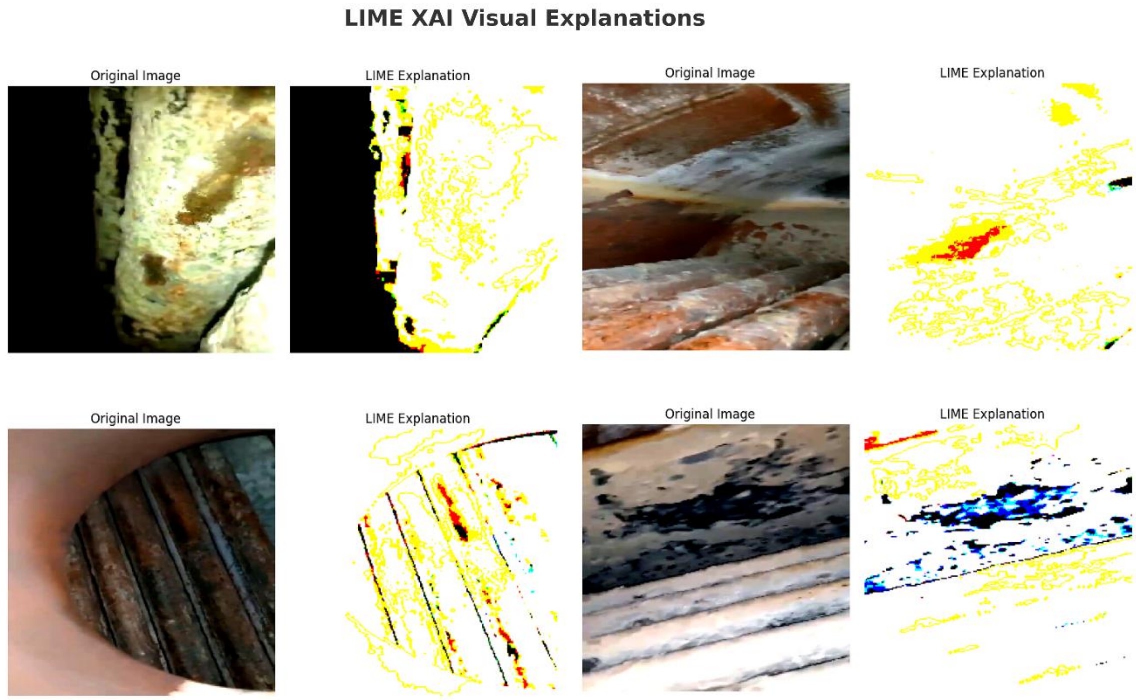


Fig 4.13: LIME Explainable AI Visual Explanations

4.4.10 Ablation Study

We performed ablation analysis to assess the effectiveness of the different modules used in the final ScaleNet V1 architecture. Multiple versions of the model were trained by susbsetting ouy-out or changing ad hoc such could be residual blocks, custom attention mechanism or dropout regularization.

- No Residual Blocks: If residual connections are not employed, the performance drops remarkably, with test accuracy decreasing nearly 6%, and loss surged. It indicates that it is the fine-grained representation which is very much propagated throughout layer depths for extracting deep features.

• No Attention Layer: The removal of attention resulted into a decrease in class-wise accuracy specifically for the minority "Others" class, which signifies the importance of attention in enhancing spatial focus and interpretability.

• No Dropout: No-dropput models were overfittning. The train accuracy was still close to 1 but the generalization gap seemed to trade off even further which indicates that dropout was necessary for regularization.

• ReLU Only vs. ReLU + Batch Normalization: The ReLU with Batch Normalization obtained faster convergence along with better convergence speed and accuracy than ReLU.

All the variants were tested under the same train-test setting. -Adding the 3-channel map and the bigger maps provided more stable results as seen with the SN V2 and SN V3 since they performed more closely to the final configuration, which kept unchanged the choice of the optimum utilized in the final version.

These results validate the necessity of the joint attention-residual-regularized hybrid design to achieve the best-in-class performance on this classification challenge.

CHAPTER 5

IMPACT ON SOCIETY, ENVIRONMENT AND SUSTAINABILITY

5.1 Impact on society

The implementation of an automated and accurate boiler scale classification system holds significant societal implications. Industrial boilers are critical components in manufacturing, power generation, and food processing sectors across Bangladesh and globally. Timely identification and categorization of scale types can reduce machine downtime, enhance operational efficiency, and minimize costly repairs. This directly contributes to economic resilience in industrial sectors and safeguards employment by preventing prolonged system failures. Additionally, by incorporating explainable AI mechanisms such as LIME, the system fosters transparency and trust among engineers and maintenance personnel, empowering them to make informed decisions based on data-driven insights.

5.2 Impact on the environment

Boiler scaling can lead to inefficient fuel consumption and excessive greenhouse gas emissions due to heat transfer inefficiencies. By accurately identifying the scale type (e.g., calcium carbonate, iron oxide, others), the proposed system facilitates targeted maintenance and preventive measures. This ensures optimal boiler performance and energy use, thereby reducing carbon footprints and limiting water and chemical wastage associated with generic cleaning procedures. Furthermore, early detection of corrosive scaling like iron oxide can prevent system leaks that might otherwise lead to the contamination of surrounding air and water resources.

5.3 Ethical Aspects

The proposed model development adheres to the ethical principles of transparency, fairness, and accountability. Data was collected ethically from real industrial settings without compromising proprietary or confidential information. The use of interpretable models through LIME explanations ensures that human stakeholders are not subjected to black-box decision-making, aligning with principles of human-in-the-loop AI. Additionally, care was taken to mitigate bias in the model, such as class imbalance handling during preprocessing and evaluation to prevent unjustified overrepresentation or neglect of rare scale types.

5.4 Sustainability Plan

To ensure long-term sustainability, the model was optimized for both computational efficiency and generalizability. This allows deployment on low-power edge devices, reducing infrastructure demands and facilitating widespread adoption across industries with varying resources. The use of modular architecture (e.g., ScaleNet V1) enables future enhancements through component updates without full retraining, promoting maintenance sustainability. Moreover, a continuous feedback loop using new real-world data and LIME-based user insights has been envisioned to support iterative model refinement, ensuring relevance in evolving industrial scenarios. Partnerships with industrial stakeholders are also being explored to integrate the system into broader predictive maintenance frameworks.

CHAPTER 6

CONCLUSION AND FUTURE WORK

6.1 Summary of the Study

This paper addressed an important and original problem in industrial process monitoring on boiler: automatic categorization of the type of boiler scales from visual data. Taking into account the operational and financial implications of the scale production at boilers, this work suggested a Hierarchical Learning Model to identify and categorise Calcium Carbonate, Iron Oxide and Miscellaneous from industrial photos. The study was based on a new labeled dataset, generated in Bangladeshi factories by domain experts. A complete methodological workflow was established with:

- Data preprocessing and augmentation for improving generalization.
- Comparison for different feature extraction methods like handcrafted features (HOG, LBP, GLCM) and deep features (VGG16, 2DSCN).
- Comparative analysis of machine learning models and Tuned XGBoost gave the best performance (90.12% test accuracy).
- Design and implementation of a custom deep learning architecture (ScaleNet V1), incorporating residual connections, attention mechanisms, and squeeze-and-excitation blocks, and optimized with Keras Tuner.

Our best network from ScaleNet architecture (we call it ScaleNet V1), has lower training loss, higher class-wise accuracy, 93.42% test accuracy and efficient inference time, compares well with modern state-of-the-art models like EfficientNet and MobileNet. An ablation study was also performed to demonstrate the efficacy of each architectural unit, we also employed LIME for explainability in the real world applicability of the model.

6.2 Conclusions

It appears to be a first application of intelligent classifying system of industrial boiler scale analysis by deep learning and machine learning which is proposed in the paper. Results On the basis of results achieved and evaluation, the proposed system was:

- Significantly better than classical ML and DL baselines.

- Computationally fast, and therefore predicted times are suitable for applications where they are used in real time.
- Explainable and trustable to help the maintenance decisionmaking applied LIME.

The effectiveness of the ScaleNet V1 verifies the possible value of adopting deep learning models in industrial inspection. It offers a scalable and automated alternative to manual and chemical methods used traditionally in blood typing. This paper is a significant advancement in the recently branched fields of smart manufacturing and predictive maintenance.

6.3 Implication for Further Study

While the current study is promising, many future directions are available:

- More Dataset: Robustness and generalization error can be improved by increasing the diversity and number of images, especially in season, light, and boiler type.
- Online Deployment: The future trajectory of this work can be towards the lightweight deployable versions of ScaleNet V1 for the onsite monitoring and not the cloud.
- Multimodal Learning — Instead of focusing on just a single modality (i.e., image data), integrating the image data with sensor readings, thermal profiles, or chemical analysis reports could potentially enable more confidence in classification or provide chances to identify any outliers that might not be seen in the image data itself.
- Domain Adaptation: An individual has separate industrial setup with few retraining, so exploring suitable segmentation and domain adaptation methods may be beneficial.

Taken together, this novel work makes a significant contribution to the field where it presents some initial directions for the development of smart boiler inspection systems, alongside a roadmap for future research at the intersection between smart maintenance, industrial automation and explainable AI.

REFERENCES

[1] R. Shohet, M. S. Kandil, and J. J. McArthur, “Machine learning algorithms for classification of boiler faults using a simulated dataset,” *IOP Conf. Ser.: Mater. Sci. Eng.*, vol. 609, no. 6, p. 062007, Sep. 2019, doi: 10.1088/1757-899X/609/6/062007.

[2] R. Shohet, M. S. Kandil, Y. Wang, and J. J. McArthur, “Fault detection for non-condensing boilers using simulated building automation system sensor data,” *Advanced Engineering Informatics*, vol. 46, p. 101176, Oct. 2020, doi: 10.1016/j.aei.2020.101176.

[3] D. Wen, Y. Pan, X. Chen, M. Aziz, Q. Zhou, and N. Li, “Analysis and prediction of thermal stress distribution on the membrane wall in the arch-fired boiler based on machine learning technology,” *Thermal Science and Engineering Progress*, vol. 28, p. 101137, Feb. 2022, doi: 10.1016/j.tsep.2021.101137.

[4] H. B. Vuthaluru, N. Kotadiya, R. Vuthaluru, and D. French, “CFD based identification of clinker formation regions in large scale utility boiler,” *Applied Thermal Engineering*, vol. 31, no. 8–9, pp. 1368–1380, Jun. 2011, doi: 10.1016/j.applthermaleng.2010.12.032.

[5] K.-B. Yoo, Y. He, H.-S. Lee, S.-Y. Bae, D.-S. Kim, and K. Shin, “Study of the scale formed on super 304H boiler tube steels after long-term steam oxidation at high temperatures,” *Materials Characterization*, vol. 146, pp. 71–80, Dec. 2018, doi: 10.1016/j.matchar.2018.09.045.

[6] J. R. Rodriguez Vasquez, R. Rivas Perez, J. Sotomayor Moriano, and J. R. Peran Gonzalez, “System identification of steam pressure in a fire-tube boiler,” *Computers & Chemical Engineering*, vol. 32, no. 12, pp. 2839–2848, Dec. 2008, doi: 10.1016/j.compchemeng.2008.01.010.

[7] M. Sohaib and J.-M. Kim, “Data Driven Leakage Detection and Classification of a Boiler Tube,” *Applied Sciences*, vol. 9, no. 12, p. 2450, Jun. 2019, doi: 10.3390/app9122450.

[8] Q. Luo *et al.*, “Automated Visual Defect Classification for Flat Steel Surface: A Survey,” *IEEE Trans. Instrum. Meas.*, vol. 69, no. 12, pp. 9329–9349, Dec. 2020, doi: 10.1109/TIM.2020.3030167.

[9] M. Prunella *et al.*, “Deep Learning for Automatic Vision-Based Recognition of Industrial Surface Defects: A Survey,” *IEEE Access*, vol. 11, pp. 43370–43423, 2023, doi: 10.1109/ACCESS.2023.3271748.

[10] S. A. Singh and K. A. Desai, “Automated surface defect detection framework using machine vision and convolutional neural networks,” *J Intell Manuf*, vol. 34, no. 4, pp. 1995–2011, Apr. 2023, doi: 10.1007/s10845-021-01878-w.

[11] Z. Chen, X. Feng, L. Liu, and Z. Jia, “Surface defect detection of industrial components based on vision,” *Scientific Reports*, 2023, doi: 10.1038/s41598-023-49359-9.

[12] R. Khanam, M. Hussain, R. Hill, and P. Allen, “A Comprehensive Review of Convolutional Neural Networks for Defect Detection in Industrial Applications,” *IEEE Access*, vol. 12, pp. 94250–94295, 2024, doi: 10.1109/ACCESS.2024.3425166.

[13] G. Prasad, G. S. Vijay, and R. Kamath C., “Comparative study on classification of machined surfaces using ML techniques applied to GLCM based image features,” *Materials Today: Proceedings*, vol. 62, pp. 1440–1445, 2022, doi: 10.1016/j.matpr.2022.01.285.

[14] B. M. Patil and V. Burkpalli, “Performance evaluation of the classifiers based on features from cotton leaf images,” *Multimed Tools Appl*, Feb. 2025, doi: 10.1007/s11042-025-20670-z.

[15] P. Simon and U. V, “Deep Learning based Feature Extraction for Texture Classification,” *Procedia Computer Science*, vol. 171, pp. 1680–1687, 2020, doi: 10.1016/j.procs.2020.04.180.

[16] S. Gudadhe, A. Thakare, and D. Oliva, “Classification of intracranial hemorrhage CT images based on texture analysis using ensemble-based machine learning algorithms: A comparative study,” *Biomedical Signal Processing and Control*, vol. 84, p. 104832, 2023, doi: 10.1016/j.bspc.2023.104832.

[17] S. Jain, G. Seth, A. Paruthi, U. Soni, and G. Kumar, “Synthetic data augmentation for surface defect detection and classification using deep learning,” *J Intell Manuf*, vol. 33, no. 4, pp. 1007–1020, Apr. 2022, doi: 10.1007/s10845-020-01710-x.

[18] J. M. Haut, M. E. Paoletti, J. Plaza, A. Plaza, and J. Li, “Visual Attention-Driven Hyperspectral Image Classification,” *IEEE Trans. Geosci. Remote Sensing*, vol. 57, no. 10, pp. 8065–8080, Oct. 2019, doi: 10.1109/TGRS.2019.2918080.

[19] P. Shamsolmoali, D. Kumar Jain, M. Zareapoor, J. Yang, and M. Afshar Alam, “High-dimensional multimedia classification using deep CNN and extended residual units,” *Multimed Tools Appl*, vol. 78, no. 17, pp. 23867–23882, Sep. 2019, doi: 10.1007/s11042-018-6146-7.

[20] N. Bai and I. Joe, “Deep learning methods with the improved attention for explainable image recognition,” *IEEE Access*, vol. 12, pp. 70559–70567, 2024, doi: 10.1109/ACCESS.2024.3397323.

[21] G. Li, Q. Fang, L. Zha, X. Gao, and N. Zheng, “HAM: Hybrid attention module in deep convolutional neural networks for image classification,” *Pattern Recognition*, vol. 129, p. 108785, Sep. 2022, doi: 10.1016/j.patcog.2022.108785.

[22] N. K. Karnam, S. R. Dubey, A. C. Turlapaty, and B. Gokaraju, “EMGHandNet: A hybrid CNN and Bi-LSTM architecture for hand activity classification using surface EMG signals,” *Biocybernetics and Biomedical Engineering*, vol. 42, no. 1, pp. 325–340, Jan. 2022, doi: 10.1016/j.bbe.2022.02.005.

[23] P. Perera and V. M. Patel, “Learning Deep Features for One-Class Classification,” *IEEE Trans. on Image Process.*, vol. 28, no. 11, pp. 5450–5463, Nov. 2019, doi: 10.1109/TIP.2019.2917862.

[24] T. Huynh-The, Q.-V. Pham, T.-V. Nguyen, T. T. Nguyen, D. B. D. Costa, and D.-S. Kim, “RanNet: Learning Residual-Attention Structure in CNNs for Automatic Modulation Classification,” *IEEE Wireless Commun. Lett.*, vol. 11, no. 6, pp. 1243–1247, Jun. 2022, doi: 10.1109/LWC.2022.3162422.

[25] I. Farady, C.-Y. Lin, and M.-C. Chang, “PreAugNet: improve data augmentation for industrial defect classification with small-scale training data,” *J Intell Manuf*, vol. 35, no. 3, pp. 1233–1246, Mar. 2024, doi: 10.1007/s10845-023-02109-0.

[26] W. Dai, D. Li, D. Tang, H. Wang, and Y. Peng, "Deep learning approach for defective spot welds classification using small and class-imbalanced datasets," *Neurocomputing*, vol. 477, pp. 46–60, 2022, doi: 10.1016/j.neucom.2022.01.004.

[27] A. Alviz-Meza, L. Hadechini-Meza, and D. Peña-Ballesteros, "Deep neural networks for external corrosion classification in industrial above-ground storage tanks," *Heliyon*, vol. 10, p. e34882, 2024, doi: 10.1016/j.heliyon.2024.e34882.

[28] L. Lin *et al.*, "Purposive Data Augmentation Strategy and Lightweight Classification Model for Small Sample Industrial Defect Dataset," *IEEE Trans. Ind. Inf.*, vol. 20, no. 9, pp. 11475–11484, Sep. 2024, doi: 10.1109/TII.2024.3404053.

[29] J. Yoo and S. Kang, "Class-Adaptive Data Augmentation for Image Classification," *IEEE Access*, vol. 11, pp. 26393–26402, 2023, doi: 10.1109/ACCESS.2023.3258179.

[30] X. Jiang and Z. Ge, "Data Augmentation Classifier for Imbalanced Fault Classification," *IEEE Trans. Automat. Sci. Eng.*, vol. 18, no. 3, pp. 1206–1217, Jul. 2021, doi: 10.1109/TASE.2020.2998467.

[31] Y. Liu, Y. Yuan, C. Balta, and J. Liu, "A Light-Weight Deep-Learning Model with Multi-Scale Features for Steel Surface Defect Classification," *Materials*, vol. 13, no. 20, p. 4629, Oct. 2020, doi: 10.3390/ma13204629.

[32] J. Yang, G. Fu, W. Zhu, Y. Cao, Y. Cao, and M. Y. Yang, "A deep learning-based surface defect inspection system using multi-scale and channel-compressed features," *IEEE Trans. Instrum. Meas.*, pp. 1–1, 2020, doi: 10.1109/TIM.2020.2986875.

[33] A. E. Houd, C. E. Hachem, and L. Painvin, "Deep Learning Model Explainability for Inspection Accuracy Improvement in the Automotive Industry," Oct. 12, 2021, *In Review*. doi: 10.21203/rs.3.rs-962895/v1.

[34] H. Sun, P. Ramuhalli, and R. E. Jacob, "Machine learning for ultrasonic nondestructive examination of welding defects: A systematic review," *Ultrasonics*, vol. 127, p. 106854, Jan. 2023, doi: 10.1016/j.ultras.2022.106854.

[35] M. Wang, "Automatic detection and classification of defects in boiler pressure vessels based on simulated annealing algorithm," *J. Phys.: Conf. Ser.*, vol. 2823, no. 1, p. 012055, Aug. 2024, doi: 10.1088/1742-6596/2823/1/012055.

[36] Z. Fei, G. West, P. Murray, and G. Dobie, "CNN-based automated approach to crack-feature detection in steam cycle components," *International Journal of Pressure Vessels and Piping*, vol. 200, p. 105112, 2023, doi: 10.1016/j.ijpvp.2023.105112.

*% detected as AI

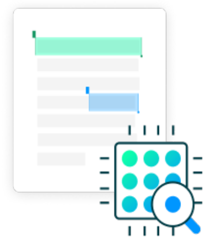
AI detection includes the possibility of false positives. Although some text in this submission is likely AI generated, scores below the 20% threshold are not surfaced because they have a higher likelihood of false positives.

Caution: Review required.

It is essential to understand the limitations of AI detection before making decisions about a student's work. We encourage you to learn more about Turnitin's AI detection capabilities before using the tool.

Frequently Asked Questions

How should I interpret Turnitin's AI writing percentage and false positives?
The percentage shown in the AI writing report is the amount of qualifying text within the submission that Turnitin's AI writing detection model determines was either likely AI-generated text from a large-language model or likely AI-generated text that was likely revised using an AI-paraphrase tool or word spinner.



False positives (incorrectly flagging human-written text as AI-generated) are a possibility in AI models.

AI detection scores under 20%, which we do not surface in new reports, have a higher likelihood of false positives. To reduce the likelihood of misinterpretation, no score or highlights are attributed and are indicated with an asterisk in the report (*%).

The AI writing percentage should not be the sole basis to determine whether misconduct has occurred. The reviewer/instructor should use the percentage as a means to start a formative conversation with their student and/or use it to examine the submitted assignment in accordance with their school's policies.

What does 'qualifying text' mean?
Our model only processes qualifying text in the form of long-form writing. Long-form writing means individual sentences contained in paragraphs that make up a longer piece of written work, such as an essay, a dissertation, or an article, etc. Qualifying text that has been determined to be likely AI-generated will be highlighted in cyan in the submission, and likely AI-generated and then likely AI-paraphrased will be highlighted purple.

Non-qualifying text, such as bullet points, annotated bibliographies, etc., will not be processed and can create disparity between the submission highlights and the percentage shown.

Md Shohidul Islam Polash Md Shohidul Islam Polash

ORIGINALITY REPORT

12%	8%	6%	6%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Daffodil International University Student Paper	2%
2	dspace.daffodilvarsity.edu.bd:8080 Internet Source	2%
3	www.frontiersin.org Internet Source	1%
4	dokumen.pub Internet Source	<1%
5	"Fifth Congress on Intelligent Systems", Springer Science and Business Media LLC, 2025 Publication	<1%
6	Submitted to University of Westminster Student Paper	<1%
7	ebin.pub Internet Source	<1%
8	dspace.bracu.ac.bd Internet Source	<1%
9	link.springer.com Internet Source	<1%
10	Arvind Dagur, Karan Singh, Pawan Singh Mehra, Dhirendra Kumar Shukla. "Intelligent Computing and Communication Techniques - Volume 1", CRC Press, 2025 Publication	<1%