People's Democratic Republic of Algeria

Ministry of Higher Education and Scientific Research

University of Kasdi Merbah Ouargla

Faculty of New Technologies of Information and Communication

Department of Computer Science and Information Technologies

Application of learning-based image representation approaches for automatic defect detection

Professional Master's Degree

Publicly defended on : 2022 /06/15

Specialty : Industrial Computer Science

Presented by : Ghilani Ala Aeddine

Siad Ismail

Supervised by : Dr .Khaldi Bilal

Jury Members

Dr. Azzaoui NADJAT	President	Ouargla Universit
Dr. AIADI OUSSAMA	Examiner	Ouargla University

Academic Year : 2021/2022

Thanks and Appreciation

I thank **ALLAH** first for being able to complete this level of study. All appreciation and respect to the generous **parents** who helped me reach this important stage in life. I thank all my **family** for the support they gave me. I thank everyone who gave me a helping hand in life, and thanks to **Dr. Khaldi Bilal**, who supervised us in completing this note. All expressions of thanks and gratitude to every **teacher** who taught me and every **student** who studied with me in my academic life.

Thank you, everyone.

ISMAIL SIAD

Thanks and Appreciation

First of all, I thank ALLAH who enabled me to reach this moment after years of hard study and work. I especially thank my parents for -their constant support during this path, and I pray to Allah that this achievement will be their pride. Thanks to the supervising professor, Dr. Khaldi Bilal, for the support and guidance during the work on this work. I thank all my classmates and all the professors for their efforts to advance scientifically and intellectually and reach the highest ranks.

GHILANI ALA AEDDINE

Contents

Abstract

The industry suffers from several problems that's causes several losses in resources and money. Most of them are due to the human element. That is why, in this project we have studied the most successful methods to find the best ways to address these problems in the field of marble industry. Where we look for deformations that affect the quality of artificial marble, which are difficult for the worker to monitor during the inspection process inside the factory by teaching the machine. As machine learning depends on the method of information processing. It provides a set of methods that facilitate the detection of deformation in the marble. Where we have done this in our work Determining the data set for the marble industry. It contains three types of defects, namely dot, crack and joint, in addition to the fourth type which is good. After that we extracted the features of those images by calculating the **STD** and the average number of pixels for all the red, green and blue **RGB** channels in our images in steps. After completing the process, we get a vector of six values while preserving the image information and the size of the storage. We get a list containing the information of all the images of the data set. Then we transformed it into a matrix to facilitate the work to get better results. Then arrange them randomly. We also used the **SKLEARN** library to split the data and calculate the accuracy of the **KNN**. Then the data was divided into two parts: 70% for training and 30% for testing. For the experiment, we choose a random image and apply the previous steps to it in addition to calculating the distance between each data set and the image we want to classify. Then we arrange the distance from the smallest to the largest and choose the first seven values, then vote and choose the most frequent category. In addition to calculating the accuracy of the results obtained, which was **66%,** which is not a good percentage. We found other ways to obtain greater accuracy. This is by using image segmentation, where it extracts image elements. Then we extracted its features and applied the **KNN** to it. And from him getting the result, which was good compared to the first result, which is **76%.**

Keywords: K-Nearest Neighbor**; K-Means ;** Image Segmentation**;** The Standard Deviation (**STD**) ; Features Extraction ; Machine Learning ; Defect Detection ; Industry Improvement.

ملخص

تعاني الصناعة من عدة مشاكل. التي تسبب عدة خسائر في الموارد والأموال . معظمهم بسبب العنصر البشري. لهذا السبب قمنا في هذا المشروع بدراسة أنجح الطرق والأساليب لإيجاد أفضل السبل لمعالجة هذه المشاكل في مجال صناعة الرخام. حيث نبحث عن التشوهات التي تؤثر على جودة الرخام والتي يصعب على العامل رصدها أثناء عملية التفتيش داخل المصنع وذلك بتعليم الالة . حيث يعتمد تعليم الآلة على طريقة معالجة المعلومات .و يقدم مجموعة من الطرق التي تسهل اكتشاف التشوه في الرخام . حيث قمنا في عملنا هذا بتحديد مجموعة البيانات الخاصة بصناعة الرخام . تحتوي على ثلاث انواع من الخلل وهي النقطة والخط والكسر بالإضافة الى النوع الرابع وهو الجيد. بعد ذلك قمنا باستخراج ميزات تلك الصور وذلك بحساب **STD** و متوسط عدد البيكسل لجميع القنوات الاحمر و الاخضر والازرق **RGB** في صورنا بخطوات .بعد الانتهاء من العملية نتحصل متجه من ستة قيم مع الحفاظ على معلومات الصور وحجم الحفظ. نحصل على قائمة تحتوي على معلومات جميع صور مجموعة البيانات .ثم قمنا بتحويلها لمصفوفة لتسهيل العمل للحصول على نتائج افضل. ثم ترتيبها بشكل عشوائي .كما استخدمنا مكتبة **SKLEARN** لتقسيم البيانات وحساب دقة ال. **KNN** ثم تم قسم البيانات الى قسمين 70٪ للتدريب و 30٪ للاختبار .وللتجربة نختار صورة عشوائية ونطبق عليها الخطوات السابقة بالإضافة الى حساب المسافة بين كل مجموعة البيانات والصورة التي نريد تصنيفها. ثم رتبانا المسافة من الاصغر الى الاكبر واخترنا اول سبعة قيم ثم نصوت ونختار الفئة الاكثر تكرارا. بالإضافة الى حساب دقة النتائج المتحصل عليها الذي كان **66**٪ وهي نسبة غير جيدة. حيث قمنا بإيجاد طرق اخرى للحصول على دقة اكبر . وذلك باستعمال **image segmentation** حيث تقوم باستخراج عناصر الصور. ثم استخرجنا ميزاتها وتطبيق عليها ال **KNN** .ومنه الحصول على النتيجة والتي كانت جيدة مقارنة بالنتيجة الاولى وهي **76**٪ .

الكلمات المفتحية :

K- الجار الاقرب , تقطيع الصورة , (STD) الانحراف المعياري , استخراج الميزات , التعلم الالي ؛ كشف عن الخلل , تحسين الصناعة.

Résume

L'industrie souffre de plusieurs problèmes. Ce qui occasionne plusieurs pertes de ressources et d'argent. La plupart d'entre eux sont dus à l'élément humain. C'est pourquoi, dans ce projet, nous avons étudié les méthodes et méthodes les plus performantes pour trouver les meilleurs moyens de résoudre ces problèmes dans le domaine de l'industrie du marbre. Où nous recherchons des déformations qui affectent la qualité du marbre artificiel, qui sont difficiles à contrôler pour l'ouvrier pendant le processus d'inspection à l'intérieur de l'usine, en apprenant à la machine. Comme l'apprentissage automatique dépend de la méthode de traitement de l'information. Il fournit un ensemble de méthodes qui facilitent la détection des déformations dans le marbre. Où nous l'avons fait dans notre travail Détermination de l'ensemble de données pour l'industrie du marbre. Il contient trois types de défauts, à savoir le point, le délaminage et la rupture, en plus du quatrième type, qui est bon. Après cela, nous avons extrait les caractéristiques de ces images en calculant le **STD** et le nombre moyen de pixels pour tous les canaux **RGB** rouge, vert et bleu de nos images par étapes. Après avoir terminé le processus, nous obtenons un vecteur de six valeurs tout en préservant les informations d'image et la taille du stockage. Nous obtenons une liste contenant les informations de toutes les images du jeu de données. Ensuite, nous l'avons transformé en matrice pour faciliter le travail et obtenir de meilleurs résultats. Disposez-les ensuite au hasard. Nous avons également utilisé la bibliothèque **SKLEARN** pour diviser les données et calculer la précision du **KNN**. Ensuite, les données ont été divisées en deux parties : 70 % pour la formation et 30 % pour les tests. Pour l'expérience, nous choisissons une image aléatoire et lui appliquons les étapes précédentes en plus de calculer la distance entre chaque ensemble de données et l'image que nous voulons classer. Ensuite, nous organisons la distance de la plus petite à la plus grande et choisissons les sept premières valeurs, puis votons et choisissons la catégorie la plus fréquente. En plus de calculer la précision des résultats obtenus, qui était de 66%, ce qui n'est pas un bon pourcentage. Nous avons trouvé d'autres moyens d'obtenir une plus grande précision. C'est en utilisant la segmentation d'image, où il extrait des éléments d'image. Ensuite, nous avons extrait ses caractéristiques et lui avons appliqué le KNN. Et de lui obtenir le résultat, qui était bon par rapport au premier résultat, qui est de **76 %**.

Mots clés : K-Plus proche voisin ; Segmentation d'images ; K-Moyennes ; L'écart type (STD) ; Extraction de fonctionnalités ; Apprentissage automatique ; détection de défauts ; Amélioration de l'industrie.

List of Figures

List of tables

General Introduction

The industry is the conversion of raw materials from one form to another. resulting a change in its use and value. This conversion may be done using natural or chemical methods or the both. These transfers may occur at a factory, small workshops, or a home. These are activities that a person performs by machine or by hand. Some of the machine relies on machine learning. It addresses the question of how to build computers that automatically improve through experience. It is one of the fastest-growing technical fields in the world today, located at the intersection of computer science and statistics, and the heart of artificial intelligence and data science. Recent advances in machine learning have been driven by the development of new learning algorithms and theories that help in the growth of many domains in the world. The industry is an essential area in the advancement of science. Where there are many industries, such as the paper industry, the cloth industry, the iron industry, the marble industry, and the wall industry...

The industry is also a phenomenon that can be studied, measured, and analysed using several criteria, such as its value, its workers, and the number of its facilities . the industry suffers from several problems the most prominent of which are the lack of some materials, the lack of their quality, or the absence of some of them. Lead to the difference in the results obtained with the requirements that we want to reach.

The marble industry is considered one of the most investment projects in the world due to its financial returns. However, it has several problems such as scratches or discoloration,. Visual inspection while making marble is an essential component of quality control. to ensure they meet quality standards. There have been many systems based on machine vision to complete this process, and in some cases, it is still performed by humans. Recent advances in machine learning have shown in this area and in many cases provide remarkable improvements in performance compared to the methods currently used to solve these problems of deformations from cracks and fractures caused during the manufacture of marble .

Computer vision is one of the solution to avoid such problems . It has several uses in different fields and facilitating human work . Computer vision is a field of computer science that aims to build smart applications capable of understanding the content of images as humans understand them. Where the image data can take many forms such as sequential

images, scenes from several cameras, data with several dimensions taken from a medical imaging device.

The objective

We aim through this work to experiment with different methods of extracting features from marble images and classifying them using machine learning techniques at the level of marble pieces by using image processing to extract the image elements with efficiency and high quality, by making sure to preserve as much information as possible for the dataset .

Chapter 1

Related works

A lot of work has been done to develop image processing processes in the field of searching for defects . The aim is to develop the industry and solve the problems it suffers from. A group of developers created and developed several algorithms that contributed to solving some problems . In this chapter, we mention a group of The work of some developers, the techniques used and the results obtained

1.1. Works related to marble industry

The industry is full of problems that need to be solved. One of them is the defect detection, which we will focus on in our research, and we will devote ourselves to talking about the field of marble industry and the methods that were used to solve this problem and compare the results that they obtained with ours.

In article [5] the experiment was carried out on natural marble images collected from a mining site in Turkey. The dataset contains 1158 images divided into 4 classes , the images was pre-processed by reduced in size from 1575*1575 to 310*310. The features extracted from it using the SDH method by extracting mean, variance, energy, correlation, entropy, contrast, and homogeneity, from the three color channels RGB. AdaBoost and SVM algorithms were used for the classification, the highest accuracy reaching 97.5 percent, by AdaBoost method .

Another study related to the detection of surface defects is entitled A Machine Learning Method for Detection of Surface Defects on Ceramic Tiles Using Convolutional Neural Networks. Which was published in 2022. Where the study was done on a data set of ceramic images containing two parts, fracture and non-fracture, with a total of 4,184 samples divided by training and testing. The data set was processed using 7 techniques to reduce data size, namely Rescaling. rotation range .shift range .shear range .flipping. Then the CNN algorithm model was taught. The accuracy of the model reached 99 percent, which is an excellent result.[6]

1.2. General related work

Researches published an article [2] entitled Automatic Fabric Defect Detection Using Learning-Based Local Textural Distributions in the Contourlet Domain. Where The fabric industry is one of the most widespread industries in the world. It is used in several products.

However, when manufacturing these products, several distortions occur that are difficult for the worker to find. Which reduces the lack of quality and prices. To find and avoid these abnormalities, the simple manual examination has been replaced by the automatic examination. To ensure high quality and in less time. Where in this article an approach is proposed to discover defects. This approach is based on three main steps. The first step is based on image analysis. The second step consists of calculating statistical signatures on labeled samples of tissue images. This step also includes BC training to distinguish between defective and non-defective fabrics. In the third and final step, the images of the checked fabrics are passed through the signature generator and BC to discover potential defects .

Another work [3] was published on defect detection called A Robust and Fast Deep Learning Based Method for Defect Classification in Steel Surfaces. It was published in 2018. They experimented with a data set called NED, which contains 1,800 images divided into 6 classes, each containing 300 samples. Before applying the model. the images were preprocessed using grayscale. They suggested using CNN deep learning methods for classification, where the accuracy of the model reached 99.96 percent, superior to ten models tested on the same dataset .

Another [4] work published an article on the study of a deep learning-based model for defect detection in laser-fusion powder layer using in-situ thermal monitoring where fabrication of metal components using laser-powder fusion is a complex process. Depends on cooling and melting layers to produce a specific part. Many processes affect the printing process, usually resulting in a set of defects. To detect these defects during printing, several techniques were proposed to monitor the processes during the cooling and melting of the layers. They used a combination of thermal imaging of the axis as a data source and a neural network based on deep learning to detect print defects. They trained the network and verified K-fold accuracy and cross-validation. The defect detection accuracy reached 96.80 percent .

Chapter 2

Methods

In this chapter, we will talk about the methods we have used to process the dataset and solve the problem that revolves around identifying the defect in the industry. And the methods that we will use are KNN first, then use KNN with image segmentation. And we explain them and how they work.

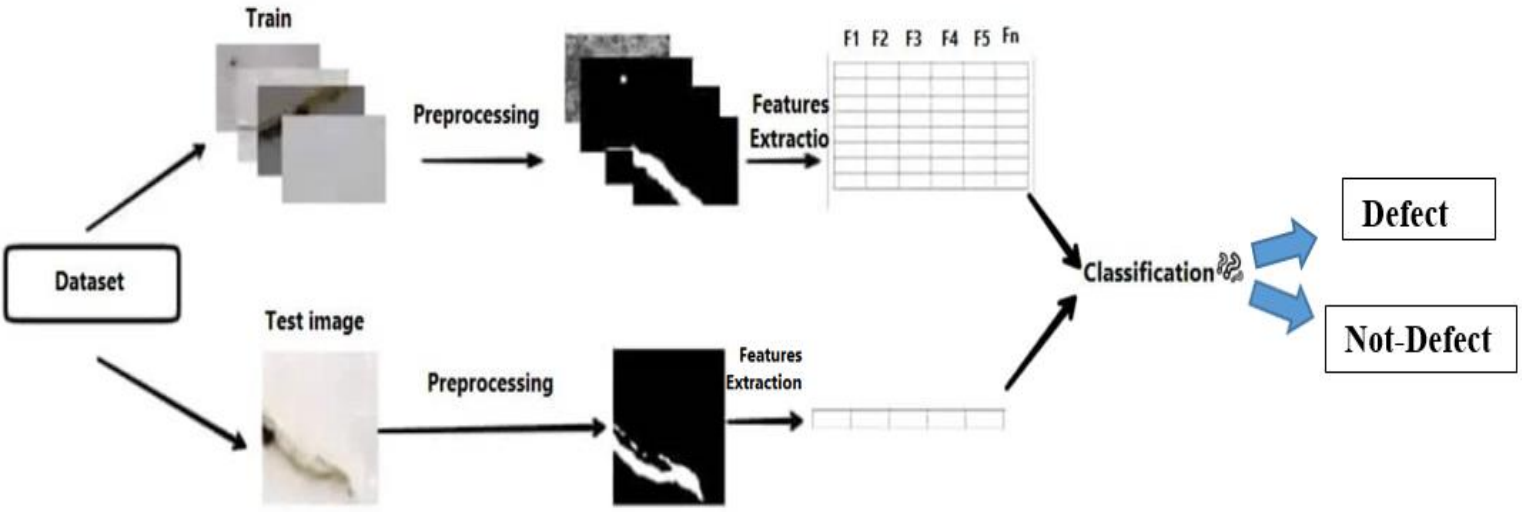


Figure 2 1:General flowchart for used method.

2.1. Pre-processing

This process is done by processing missing values and removing unwanted or annoying data. It is intended to improve image data that prevents unwanted distortions or enhances certain image features relevant to an additional processing and analysis task.

2.1.1. Image Segmentation

Image segmentation is a method in which a digital image is divided into different subgroups called image segments, which helps reduce image complexity for simpler image analysis and further processing. Division in easy words is to assign labels to pixels. All image elements or pixels of the same class have a common label assigned to them. For example: let's take a problem where the image must be provided as input for object detection. Instead of processing the entire image, the detector can be fed into a defined area by the hashing

algorithm. This will prevent the detector from processing the entire image and thus reduce the inference time . Image segmentation techniques :

.

- Segmentation based on threshold.
- Edge segmentation.
- Division by region.
- Hash-based aggregation.
- Segmentation based on artificial neural network.
- Image Segmentation Using K –means [1].

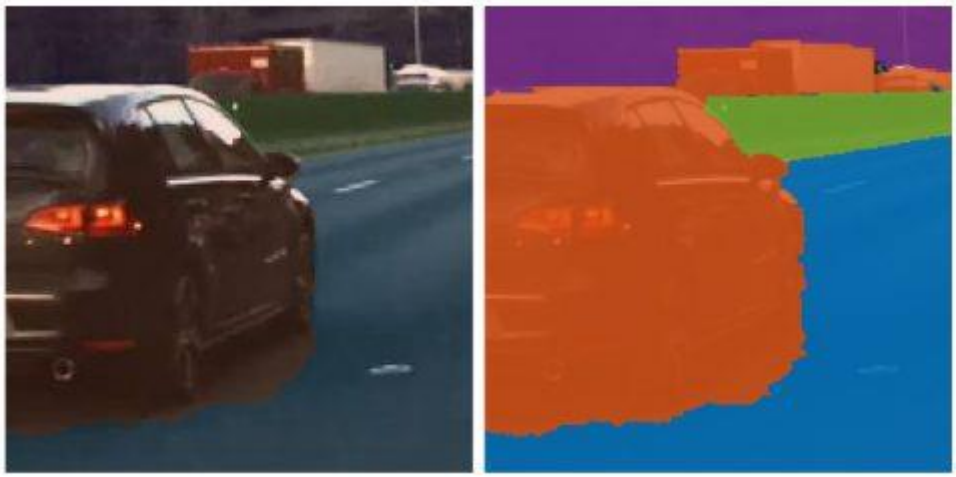


Figure 2 2: Image segmentation example .

In this article, we will use Threshold Based and Image Segmentation Using K -means.

Segmentation based on threshold :

Thresholding is a common segmentation technique used to separate an object from its background. Thresholding involves comparing the value of each image pixel to a given boundary. This divides all the pixels of the input image into two groups:

1. Pixels have a density greater than the minimum.
2. Pixels have a density less than the minimum.

The formula of the Thresholding method is : $p_i=f_i/N$ $\sum_{i=1}^{L} p_i=1$ $p_i \geq 0$.

Image Segmentation Using K –means :

Clustering is a method to divide a set of data into a specific number of groups. It's one of the popular method is k-means clustering. In k-means clustering, it partitions a collection of data into a k number group of data. It classifies a given set of data into k number of disjoint cluster. K -means algorithm consists of two separate phases. In the first phase it calculates the k centroid and in the second phase it takes each point to the cluster which has nearest centroid from the respective data point. There are different methods to define the distance of the nearest centroid and one of the most used methods is Euclidean distance. Once the grouping is done it recalculate the new centroid of each cluster and based on that centroid, a new Euclidean distance is calculated between each center and each data point and assigns the points in the cluster which have minimum Euclidean distance. Each cluster in the partition is defined by its member objects and by its centroid. The centroid for each cluster is the point to which the sum of distances from all the objects in that cluster is minimized. So K -means is an iterative algorithm in which it minimizes the sum of distances from each object to its cluster centroid, over all clusters .

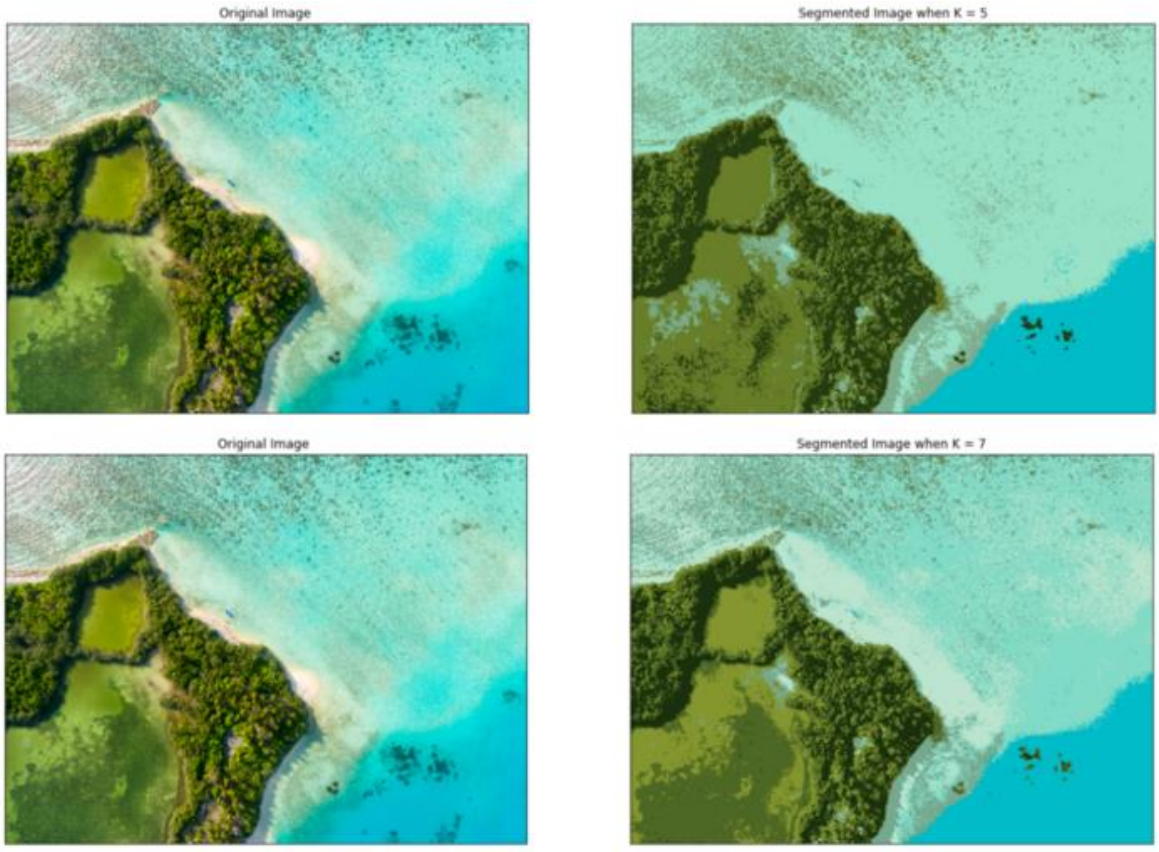


Figure 2 3: Image Segmentation using K-Means .

2.2.Features Extraction

In real life, all the data we collect are in large amounts. To understand this data, we need a process. Manually, it is not possible to process them. Here's when the concept of feature extraction comes in.

Feature extraction is a part of the dimensionality reduction process, in which, an initial set of the raw data is divided and reduced to more manageable groups. So when you want to process it will be easier. The most important characteristic of these large data sets is that they have a large number of variables. These variables require a lot of computing resources to process. So Feature extraction helps to get the best feature from those big data sets by selecting and combining variables into features, thus, effectively reducing the amount of data. These features are easy to process, but still able to describe the actual data set with accuracy and originality . The features extraction has same benefits which is useful when you have a large data set and need to reduce the number of resources without losing any important or relevant information. Feature extraction helps to reduce the amount of redundant data from the data set .

2.2.1. Applying Feature Extraction

To extract the features from our dataset images, first, we calculated the STD and the mean number of pixels for all the red, green, and blue (RGB) channels in our image, according to the following steps:

Mean calculation

Calculate the colors value of the channel and divide it by the total number of pixels

Mean (RGB) $= \frac{(\sum_{n}^{i=0} p(rgb))}{x \times y}$

P: the pixel value.in the 3 channels

N: is the number of pixels in the image.

X$\times$ y: the number of all the pixels, which is(256×256).

STD calculation

The standard deviation (STD) is a useful measure of the spread of normal distributions. In normal distributions, the data is distributed symmetrically without skew. Most values are grouped around a central region, with values decreasing as you move away from the center. The standard deviation tells you how to spread from the center of the distribution where your data is on average .

At first, we calculate deviations by using the following equation :

$Deviations = \sum_{n}^{i=0}(xi - m)^2$

m: is the mean number of the channel pixel.

$Variance = \frac{Deviations}{n}$.

$STD = \sqrt{Variance}$.

We apply this process to the three-color channels.

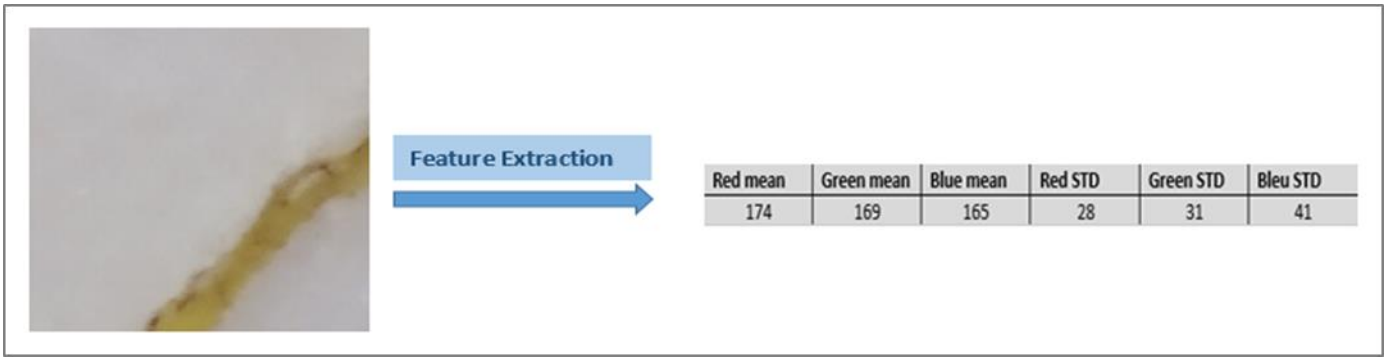


Figure 2 4 :Example of Feature extraction from image.

2.3. Classification

The Classification algorithm is a Supervised Learning technique that is used to identify the category of new observations on the basis of training data. In Classification, a program learns from the given dataset or observations and then classifies new observation into a number of classes or groups. Such as, Yes or No, 0 or 1, Defect or Not Defect, cat or dog, etc. Classes can be called as targets/labels or categories.

The algorithm which implements the classification on a dataset is known as a classifier. There are two types of Classifications . **Binary Classifier** for the classification problem has only two possible outcomes. **Multi-class Classifier** for classification problem that has more than two outcomes , and some of the best classification algorithms are K-Nearest Neighbours, Support Vector Machines , Naïve Bayes , Decision Tree Classification , Random Forest Classification and Logistic Regression [7].

And for our experiment we only using the KNN algorithm .

2.3.1. K- Nearest Neighbors (KNN) :

A k-nearest-neighbors algorithm, often abbreviated KNN, is an approach to data classification that estimates how likely a data point is to be a member of one group or the other depending on what group the data points nearest to it are in , K-NN An algorithm, looking at one point on a grid, trying to determine if a point is in group A or B, looks at the states of the points that are near it. The range is arbitrarily determined, but the point is to take a sample of the data. If the majority of the points are in group A, then it is likely that

the data point in question will be A rather than B, and vice versa. The k-nearest-neighbors is an example of a "lazy learner" algorithm because it does not generate a model of the data set beforehand. The only calculations it makes are when it is asked to poll the data point's neighbors. This makes k-NN very easy to implement for data mining . K-NN Properties:

- K-Nearest Neighbor is one of the simplest Machine Learning algorithms based on Supervised Learning technique.
- K-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories.
- K-NN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category by using K- NN algorithm.
- K-NN algorithm can be used for Regression as well as for Classification but mostly it is used for the Classification problems.
- K-NN is a **non-parametric algorithm**, which means it does not make any assumption on underlying data.
- It is also called a **lazy learner algorithm** because it does not learn from the training set immediately instead it stores the dataset and at the time of classification, it performs an action on the dataset.
- KNN algorithm at the training phase just stores the dataset and when it gets new data, then it classifies that data into a category that is much similar to the new data .

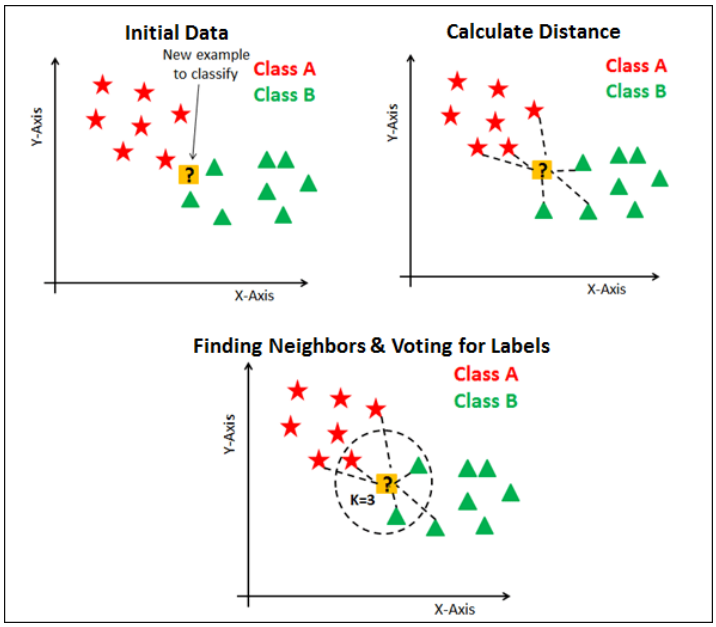


Figure 2 5 : K-NN representation [8].

There are many ways to calculate the distance in KNN, for example :

- The Manhattan Method :

  p_1 at (x_1, y_1) and p_2 at (x_2, y_2), it is $|x_1 - x_2| + |y_1 - y_2|$.

- Euclidean Method :

  $d(x\ ,\ y)=\sqrt{\sum_{i=1}^{n}(yi - xi)^2}$.

Chapter 3

Experiment and Discussion

In this chapter, we will apply the proposed machine learning methods in the previous chapter. Defining the hardware and software used in the application, in addition to presenting the data set in detail and presenting and discussing the results.

3.1.Development Tools

The hardware environment . In our experiment from feature extraction to the methods apply, we used two devices with the following characteristics on all models:

- DELL Pc , windows 8.1 Enterprise with 4,00 GB memory capacity, processor Intel(R)Core(TM) i3-6006U CPU @2.00 GHz, and system 64 bits.
- TOCHIBA Pc , windows 10 pro with 4,00 GB memory capacity, processor Intel(R)Pentium(R) CPU N3540 @2.16 GHz, and system 64 bits.

The software environment :

Table3. 1: The software environment.

Environment.	Description
programming language	Python 3.10
Python Library	Numpy , matplotlib ,cv2 ,OS , sklearn
Distribution	pycharm

3.2.Accuracy

Accuracy is one metric for evaluating classification models. Informally, **accuracy** is the fraction of predictions our model got right. Formally accuracy has the following definition :

$$\boldsymbol{Accuracy = \frac{Number\ of\ correct\ predictions}{Total\ number\ of\ predictions}}$$

3.3.The dataset

The dataset we used in our search was downloaded from web site kaggle.com by the name (marble surface anomaly detection 2)[9] dataset. It consists of 2,249 images of the marble surface taken with a smartphone camera in jpg format. The images are in color RGB have dimensions of 256 * 256. Labeled into four classes (dot, crack, and joint) and another class is the good. And each class contain (crack = 984 image ,dot = 92 image , good = 860 image and joint = 313 image). Each image from the dataset contains only one type of defect.

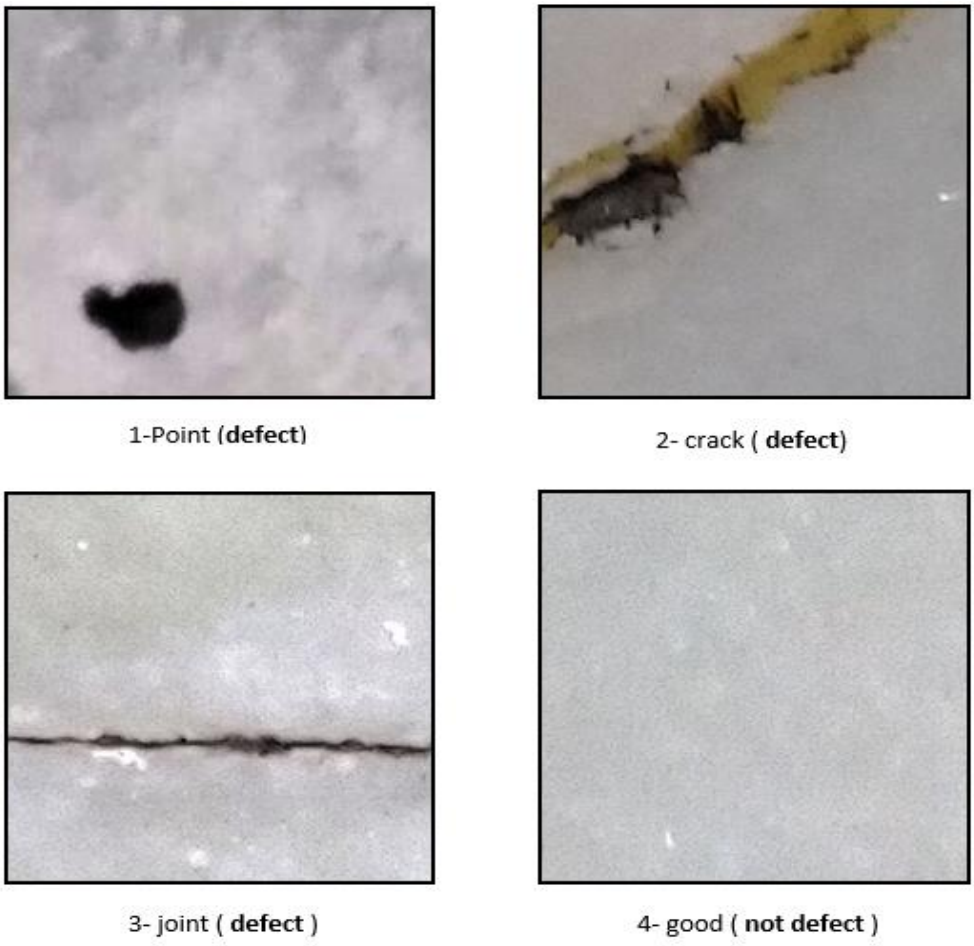


Figure 3 1: Sample of Marble Surface Classify DenseNet201 with classification.

In this experiment we will use 2 scenarios. In the first scenario, we will use the original images without pre-processing and extract the features from them by calculating the STD and MEAN of the RGB channels and then classifying them. In the second scenario, we process the images, then divide the images into parts (4..16), extract the features from the images, and then classify them.

3.4.Scenario 1

- We calculate the STD and the Mean on the three color channels RGB for all images on the data set. We get a vector of six values from each image . and one more for the label.
- After that we get a list containing the information of all the images from the dataset.
- Then convert the list of features to an array to make it easier to work with .
- To obtain better experimental results, we randomly shuffle the full data.
- Then split the data into two parts, the first is for training which contains 70% of the total data . The second section contains 30% for the test.
- After that the KNN algorithm well predict the classification of the test data .

3.4.1. The Result of KNN

The results we obtained after trying to classify a group of images using KNN.

3.4.2. Accuracy after applying KNN

After we apply the KNN algorithm to a set of images in the dataset, we get percentages of accuracy = 66% was the largest percentage in the k optimal = 20 . And the following [Figure 3 2]curve shows the change in accuracy of the model in terms of K .

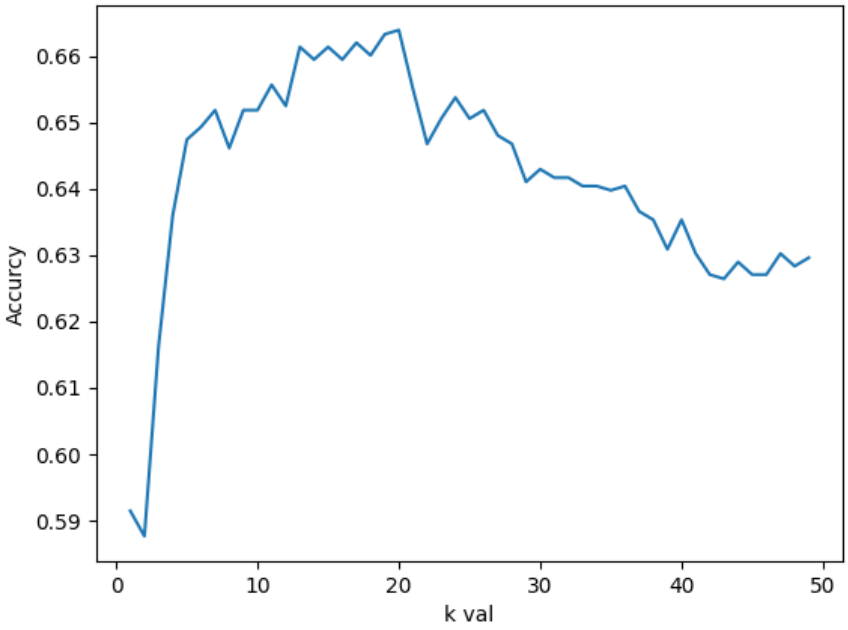


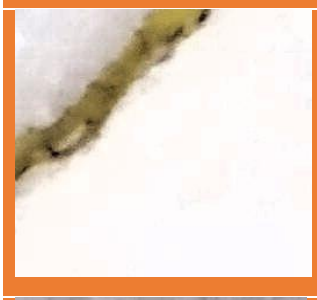
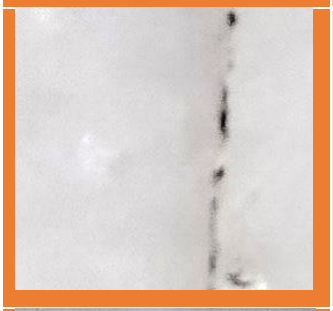
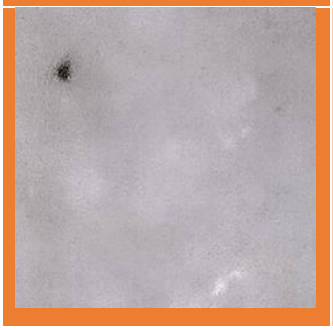
Figure 3 2: plot of the KNN model Accuracy .

After obtaining an accuracy rate of 66 percent, we find that it is not good enough. We assume that the reason is due to the inefficiency of the features extracted. A lot of image information was lost and this negatively affected the result. We can improve it by using other effective methods of extracting features and image processing, and this is what we tried to do in the following application .

Example of our KNN algorithm

As a practical experiment to test the model. We choose a random image from the data set and apply the same steps to extract the features mentioned earlier. Convert it to a list. Then we calculate the distance between each dataset and the image that we want to classify using the Euclidean distance method., we chose the K=7 neighbors. Then we vote and choose the most frequent category.

Table3. 2:Test our KNN model.

The images	7-nearest neighbours	Classification	true or false
	['crack','crack', 'crack','crack', 'crack', 'good', 'joint']	Crack	True
	['crack','good', 'good', 'good', 'good', 'joint', 'crack']	Good	False
	['good', 'good', 'good', 'crack', 'joint', 'good', 'good']	Good	True
	['crack', 'joint', 'joint', 'good', 'good', 'crack', 'joint']	Joint	True
	['crack', 'dot', 'dot', 'dot', 'good', 'good', 'dot']	Dot	True

We note from the table that the model was correct in classifying 4 samples out of the 5 images, and it erred in classifying the DOT defect. The reason is that the feature vector of the image to be classified has more values close to the values of the class GOOD, and this is the main reason for the poor accuracy of the model.

3.5.Scenario 2

In second experiment we preprocessed the dataset by using image segmentation and the image split technique then extract the features .

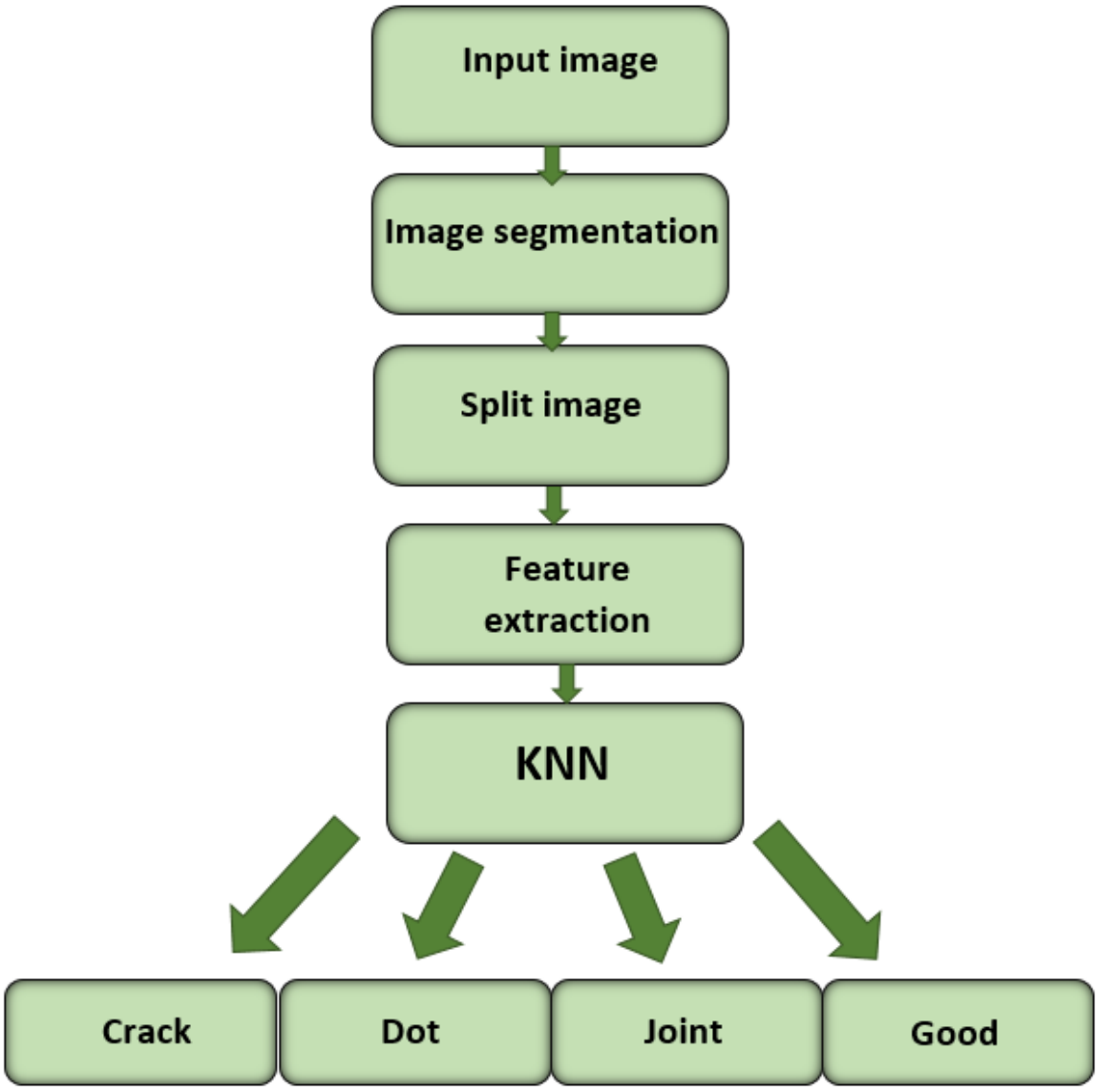


Figure 3 3 : the Scheme shows the steps involved in Scenario 2.

Image Pre-processing :

After completing the first experiment which is **KNN**. We decided to improve feature extraction from dataset images using the image segmentation method in order to process the images.

- In this step we will focus on local features.
- The image segmentation method divides the content of each image into two parts defect and non-defect part.
- In order to do this We use the K-Means segmentation method to segment the image into two parts, by setting the value of K = 2 in this case. This technique searches for nearby pixels and assigns them the same color. We get an image of two different colors.
- After that we apply the threshold to it . And that to get a batter binary segmentation result .

The following image shows the steps for image segmentation we used :

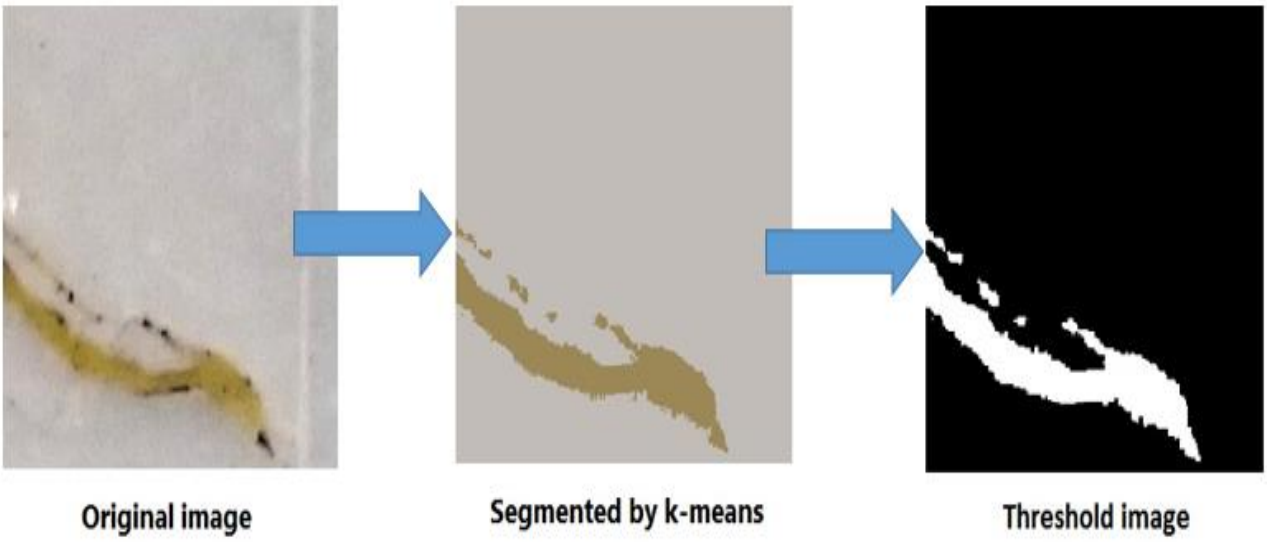


Figure 3 4: Image segmentation steps

splitting images:

In the second stage, we will start the feature extraction stage. To do this process, we searched by the best possible way to locate the defect in the image and extract its information . The solution we came up with it is to divide the image into 4 parts (top right, top left, bottom right, bottom left). This method enabled us to determine the location of the defect in the parts of the image and how much space it takes in the marble piece. To determine the common pattern between each type of defect.

The dot defect can be easily identified because in most cases it is located in only one place of the image parts. The dot also takes up less space than the rest of the defects and its shape is close to the circle in most cases.

The crack defect can be determined by the fragment images. We notice that the crack is distributed over more than one image , and This is done by calculating the number of white pixels in each part of the images. It also takes a large area on the marble piece compared to the dot defect ,and its shape is tilted , and by calculating how it is distributed in the parts of the image.

The joint defect can determined by how it is distributed on the images, as it extends over more than one part of the image, such as the crack defect, but it takes up less space than it, and It is distributed straight across compared to the crack defect .

The good class well determined by the absence of a specific form of defect , so it can be easily distinguished from other defects.

After doing this procedure, we extracted the features for each image separately. Where we calculate the number of black and white pixels for each image. Then we calculate the STD and mean number of pixels for each part of the 4 images part . With this, we get a list Contains 10 values form each image in the data set, as shown in the following figure.

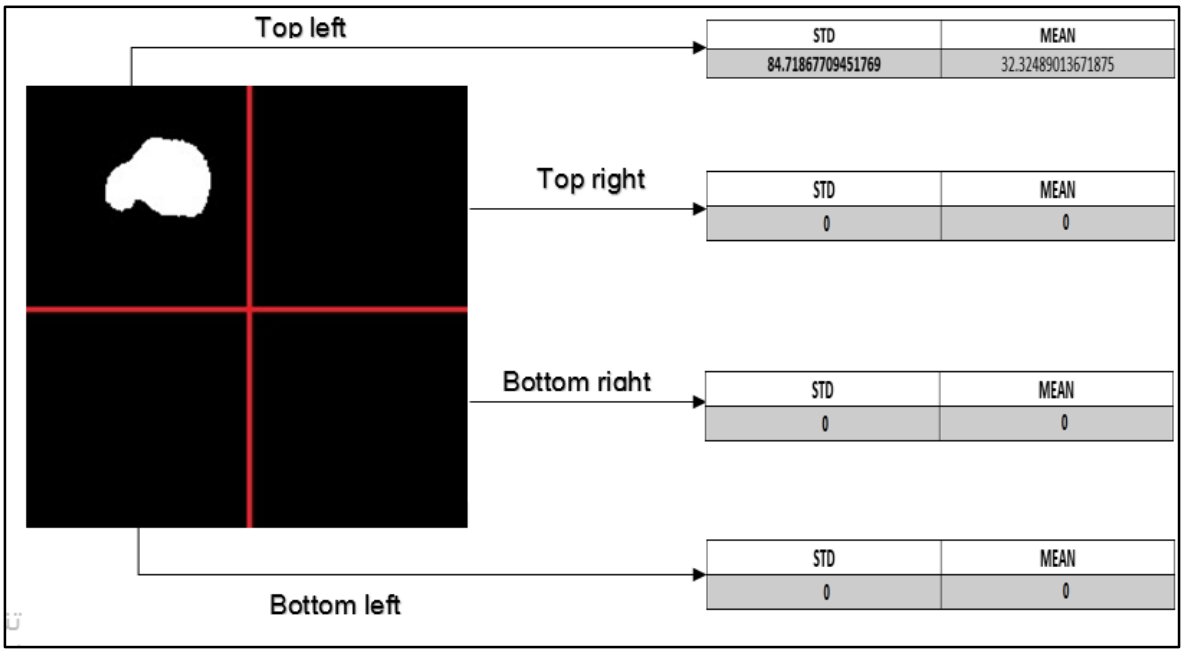


Figure 3 5:Steps for Feature extraction from segmented images .

Result of applying kNN on segmented images

To calculate the average accuracy of the model, we experimented with random experiment of the data and then calculated the mean. The average accuracy reached 75%. The following table shows the experimental cases.

Table3. 3:Some examples of accurate segmentation images.

Random state	K optimal	accuracy
1	4	0.7459043574967142
3	12	0.7464988373268628
6	9	0.7484015771913861
9	3	0.7312445657668588
12	14	0.7509614801334547
15	12	0.7579577393590132
21	7	0.7471539783641694
23	7	0.7458618946517036
25	6	0.7407906177332929
27	18	0.748411687392579
31	30	0.7528884844808412
35	10	0.7446183399049642
37	44	0.7401516530178951
39	8	0.7592134263471844
40	14	0.7522394095642504
42	11	0.7458881811748055
46	5	0.7604670912951168
47	21	0.7515964007683753
48	18	0.7426974016782933
49	23	0.7465150136487717

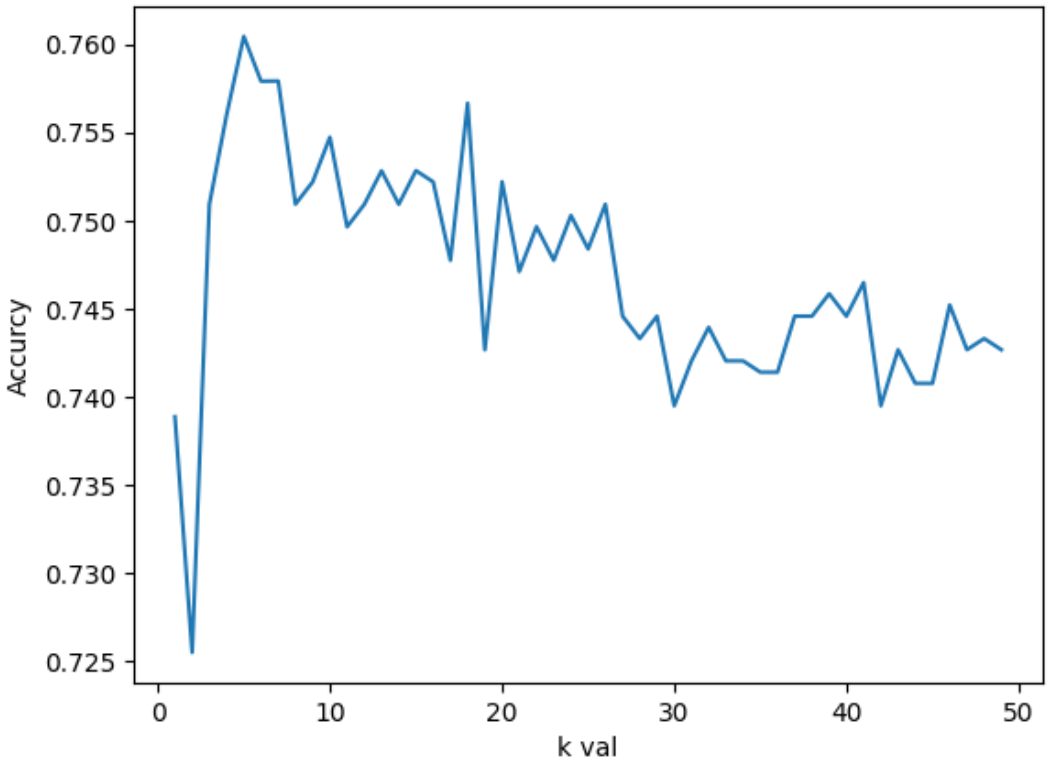


Figure 3 6 :Curve of Images Segmentation and KNN Accuracy .

We note that after applying the KNN to the features extracted from splited images , the accuracy of the model increased by 10 % compared to the previous model, where it reached 75 %. This is due to the quality of the preprocessing provided by segmentation images. And splitting images.

We also tried dividing the image into more than 4 parts. To 16 parts in an effort to extract as much information from the images. And we noticed an increase in the image processing time, which means an increase in resource consumption and the stability of accuracy in the range of 76 %. That is why we have adopted the division into 4 only

General Conclusion

The industry field suffers from several problems that affect the quality of products. One of them is the presence of distortions during their manufacture that require human effort and a large number of workers to monitor them. That is why we aim through our project to solve this problem by replacing the human examination with a machine. Because it is less expensive and more effective and increases the speed of work. Where we chose the field of marble industry for study. We focused in our project on finding the most important means of automatic detection of defects in images. Using different methods of machine learning while calculating the accuracy of the model in the detection process. In Scenario 1, we extract the STD and MEAN features and then classify using KNN on the images. Where the accuracy of the model in discovering defects reached 65 percent, which is an ineffective percentage to rely on in the field. We assume that the reason for the poor accuracy is due to the loss of a lot of data during the feature extraction process. To improve this result, in Scenario 2, we pre-process the images using image segmentation to extract the local features of the images. Then divide the images into four parts to extract the largest number of features. We carry out the classification process using KNN Algorithm. The accuracy of the model reached 75 percent is good compared to the first translation. The reason is due to the pre-processing of the images, where we were able to reach the largest number of features compared to the first scenario.. The result we reached is that the pre-processing and extraction Features have a significant role in improving the quality of classification. We suggest to improve the results obtained by using newer techniques for identifying features such as bag of visual words + (SIFT or SURF or ORB or BRISK) . Other classification algorithms such as SVM or deep learning can be used CNN.

Bibliography

[1] Yanan Guoa , Zengshun Zhao* and Zuohong Wub Shandong University of Science and Technology, China *Corresponding author " Research on Image Segmentation based on Full Convolutional Neural Network ". In 2020 .

[2] Daniel Yapi , Mohand Saïd Allili, Member, IEEE, and Nadia Baaziz, Member, IEEE , " Automatic Fabric Defect Detection Using Learning-Based Local Textural Distributions in the Contourlet Domain", IEEE TRANSACTIONS ON AUTOMATION SCIENCE AND ENGINEERING, VOL. 15, NO. 3, JULY 2018.

[3] Fatima A .Saiz , Ismael Serramo , Inigo Barandiaran , jairo R Sanchez ," A Robust and Fast Deep Learning-Based Method for Defect Classification in Steel Surfaces ".In 2018.

[4] Hermann Baumgartl, Josef Tomas, Ricardo Buettner, and Markus Merkel ," A deep learning-based model for defect detection in laser-powder bed fusion using in-situ thermographic monitoring ".In 4 September 2019 .

[5] Dokuz Eylül University, Department of Electrical and Electronics Engineering, Kaynaklar Campus published an article called ,"Using AdaBoost classifiers in a hierarchical framework for classifying surface images of marble slabs", "Azmir Turkey ",in 2020.

[6] Okeke Stephen , Uchenna Joseph Maduh , Mangal Sain ," A Machine Learning Method for Detection of Surface Defects on Ceramic Tiles Using Convolutional Neural Networks".

[7] Pratap Chandra Sen, Mahimarnab Hajra and Mitadru Ghosh , "Supervised Classification Algorithms in Machine Learning: A Survey and Review " , Emerging Technology in Modelling and Graphics, Advances in Intelligent Systems and Computing 937 , in 2020 .

[8] https://towardsdatascience.com/machine-learning-basics-with-the-k-nearest-neighbors-algorithm-6a6e71d01761 in 22/06/2022.

[9] the dataset link https://www.kaggle.com/datasets/wardaddy24/marble-surface-anomaly-detection-2 in 22/06/2022 .