International Journal of Applied Research and Sustainable Sciences (IJARSS)
Vol. 2 No. 6, 2024: 417 - 426

Comparison of Defective Casting Product Classification Results Using the K-Nearest Neighbors Algorithm

Muhammad Farhan Alfarizi[1], Muhamad Fatchan[2], Wahyu Hadikristanto[3]
Universitas Pelita Bangsa
Corresponding Author: Muhammad Farhan Alfarizi, farhan.19@mhs.pelitabangsa.ac.id

A R T I C L E I N F O

Keywords: Product, Casting, Prediction, KNN, Naïve Bayes

Received : 7 April
Revised : 8 May
Accepted: 9 June

A B S T R A C T

This study compares the accuracy of K-Nearest Neighbors (KNN) and Naive Bayes algorithms in detecting defects in impeller products. Using a dataset of impeller images, we applied preprocessing, feature extraction, and selection techniques. The models were assessed using metrics such as precision, accuracy, F1-score, recall. and with KNN achieving 98.11% accuracy and Naive Bayes 85.38%. The t-SNE visualization confirmed distinct clustering of defective and non-defective products. Our findings suggest that KNN is more reliable for defect detection in industrial applications. These results provide valuable insights for implementing effective machine learning models in manufacturing quality control.

INTRODUCTION

The manufacturing industry worldwide, including in Indonesia, continues to experience rapid development with the adoption of advanced technologies to enhance quality and production efficiency. One of the main challenges in this industry is detecting defects in casting products, such as porosity, cracks, and inclusions. Quick and accurate defect detection is crucial to ensure product quality and reduce production costs due to defects(Syahril Dwi Prasetyo et al., 2023).

Traditionally, defect inspection has been performed through visual examination by human operators, which heavily relies on individual skills and experience. However, this method has several limitations, including subjectivity, fatigue, and the inability to process large volumes of data quickly. Therefore, there is an urgent need to develop more reliable and efficient automated methods(Karyadi, 2023).

In recent years, advancements in machine learning have offered potential solutions to this problem. Algorithms such as K-Nearest Neighbors (KNN) and Naive Bayes have shown promising results in defect detection applications. KNN is a simple yet a highly efficient algorithm that categorizes data based on its proximity to training data. Conversely, Naive Bayes is a classification algorithm that relies on Bayes' Theorem, assuming that features are independent of each other (Zakiyah et al., 2022).

This study aims to compare the accuracy of defect detection in casting products using the KNN and Naive Bayes algorithms. The dataset used consists of images of impeller products to be installed in a centrifugal pump. This The dataset is divided into training and testing sets to evaluate the performance of each model. The evaluation metrics used are precision, accuracy, recall, and F1-score. (Rinanda et al., 2022).

This study is expected to offer insights into the most effective machine learning algorithms for defect detection in casting products, thereby improving the efficiency and quality of manufacturing processes. The findings of this research can also serve as a basis for further development in the implementation of automated defect detection systems in the industry.

LITERATURE REVIEW

K-Nearest Neighbors (KNN) Theory

K-Nearest Neighbors (KNN) is one of the most commonly used machine learning algorithms for classification tasks. This algorithm classifies new data based on the proximity or distance to the existing training data. KNN is known for its simplicity and its ability to handle data with various distributions. However, the performance of KNN is highly dependent on the selection of the parameter k (Sitepu & Manohar, 2022).

Previous studies have shown that KNN can be effectively used for defect detection in casting products. Despite its advantages, KNN has limitations in terms of scalability and speed when handling large and imbalanced datasets. Therefore, parameter optimization and pre-processing techniques are required to enhance the performance of KNN in this context (Sitepu & Manohar, 2022).

H1: K-Nearest Neighbors has higher accuracy in detecting defects in impeller products that will be installed on centrifugal pumps compared to Naive Bayes.

Naive Bayes Theory

Naive Bayes is a classification algorithm based on Bayes' Theorem, assuming that the features are mutually independent. Even though this assumption is rarely true in real-world applications, Naive Bayes frequently delivers strong performance due to its simplicity and efficiency in handling data. This algorithm is suitable for classification tasks involving many features, such as defect detection in casting products (Cahyo, 2023). Previous research has shown that the Naive Bayes algorithm has limitations in the accuracy of defect detection in casting products. For instance, a study by (Nurwijayanti, 2023). Indicated that Naive Bayes produced less satisfactory accuracy compared to other algorithms like KNN. Therefore, this study uses KNN to improve the accuracy of defect detection, leveraging its strength in modelling complex non-linear relationships.

H2: Naive Bayes has lower accuracy in detecting defects in casting products compared to K-Nearest Neighbors (Kaharudin et al., 2023).

METHODOLOGY

This study utilizes a quantitative approach with experimental methods to detect defects in casting products. It employs two machine learning algorithms, K-Nearest Neighbors (KNN) and Naive Bayes, to classify images of impeller products intended for installation in centrifugal pumps.

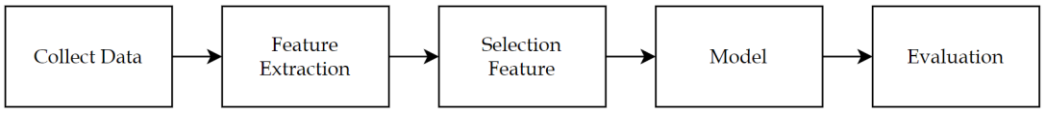


Figure 1. Conceptual Framework

Images of casting products were collected from Kaggle and used as a dataset for training and testing. This dataset consists of two categories: NG (defective) products and OK products (*Casting Product Image Data for Quality Inspection*, n.d.).

Collect Data

The sample used includes images of impeller products obtained from the Kaggle dataset (*Casting Product Image Data for Quality Inspection*, n.d.). This dataset comprises images of products that have been labelled as defective (NG) and non-defective (OK). The images are 300x300 pixels in size with an average file size of approximately 9681 KB. After processing as needed, this data is used with 5-fold cross-validation to determine the best method for detecting defects in casting products.

Feature Extraction

In this study, images of impeller products were first converted to grayscale and resized to 300x300 pixels to standardize the data (Liu et al., 2023). Histogram of Oriented Gradients (HOG) and Local Binary Patterns (LBP) were then applied to capture edge, gradient, and texture information, respectively (Liu et al., 2023). Principal Component Analysis (PCA) was employed to reduce dimensionality, preserving the most significant features while minimizing computational load. Finally, the features were normalized to ensure comparability, enhancing the performance and stability of the KNN and Naive Bayes models in defect detection (Jin & Chen, 2022).

Selection Feature

Feature selection in this research aimed to determine the most relevant features for improving the predictive power of machine learning models. A correlation matrix was used to eliminate highly correlated and redundant features, reducing model complexity. Recursive Feature Elimination (RFE) was applied to iteratively rank and remove the least important features, ensuring the retention of significant ones. Additionally, feature importance scores from the Random Forest algorithm were used to identify and select the most influential features. These techniques enhanced the efficiency and accuracy of the KNN and Naive Bayes models in detecting defects in impeller products (Darst et al., 2018).

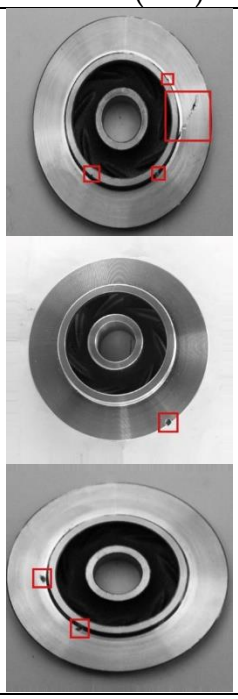
Image Sample OK	Image Sampel Defect (NG)
	

Table 1. Selection Feature

Model

K-Nearest Neighbors (KNN) and Naive Bayes are two machine learning algorithms used in the model training and evaluation process. KNN is a simple, non-parametric technique that uses cross-validation to discover the ideal k parameter before classifying data points depending on who their nearest neighbors are. Based on Bayes' Theorem, the probabilistic classification algorithm Naive Bayes was trained using the Gaussian method and assumes feature independence. Metrics like precision, recall, precision, the F1 score, Matthews Correlation Coefficient (MCC), and Area Under the Curve (AUC) were used to evaluate both models. To guarantee robustness and dependability, a 5-fold cross-validation was used. The goal was to create precise and effective models for impeller product defect identification that would ensure their usefulness in practical applications (Sheth et al., 2022).

Evaluation

This study evaluated the efficacy of the Naive Bayes and K-Nearest Neighbors (KNN) models in detecting impeller product faults. Several measures were used in the evaluation, including F1-score, recall, accuracy, and precision.

Accuracy is defined as the proportion of correctly classified instances among the total instances:

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

Where TP is True Positive, TN is True Negatives, FP is False Positive, and FN, is False Negatives.

Precision is the ratio of true positive predictions to the total predicted positives:

$$Precision = \frac{TP}{TP + FP}$$

Recall (also called Sensitivity) measures the proportion of actual positives correctly identified

$$Recall = \frac{TP}{TP + FN}$$

F1-score is the harmonic mean of precision and recall

$$F1 - score = 2\ x\ \frac{Precision\ x\ Recall}{Precision + Recall}$$

Matthews Correlation Coefficient (MCC) provides a balanced measure for binary classifications, taking into account true and false positives and negatives

$$MCC = \frac{(TP\ x\ TN) - (FPxFN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

A 5-fold cross-validation procedure was used to guarantee robustness. In order to do this, the dataset is divided into five subgroups. The models are then repeatedly trained and tested using various combinations of these subsets. This approach lowers the chance of overfitting and provides a more accurate assessment of the model's performance.

RESEARCH RESULT

The effectiveness of the K-Nearest Neighbors (KNN) and Naive Bayes models in detecting defects in impeller products was assessed through a performance comparison. Metrics like accuracy, F1-score, precision, and recall were used in the evaluation.

Table 2. Evaluation of Comparative Results

Dataset	Model	AUC	CA	F1	Precision	Recall	MCC	Accuracy
Casting Product	KNN	0.99	0.98	0.98	0.98	0.98	0.96	98.11%
	Naïve Bayes	0.86	0.85	0.85	0.85	0.85	0.72	85.38%

The Naive Bayes model, while demonstrating decent performance with an overall accuracy of 85.38%, showed lower precision and recall compared to the KNN model. This suggests that KNN is more efficient in accurately classifying both defective and non-defective impeller products. The KNN model demonstrated high performance with an overall accuracy of 98.11%. It achieved perfect precision for defective products and excellent recall for non-defective products, indicating its reliability in identifying both classes accurately.

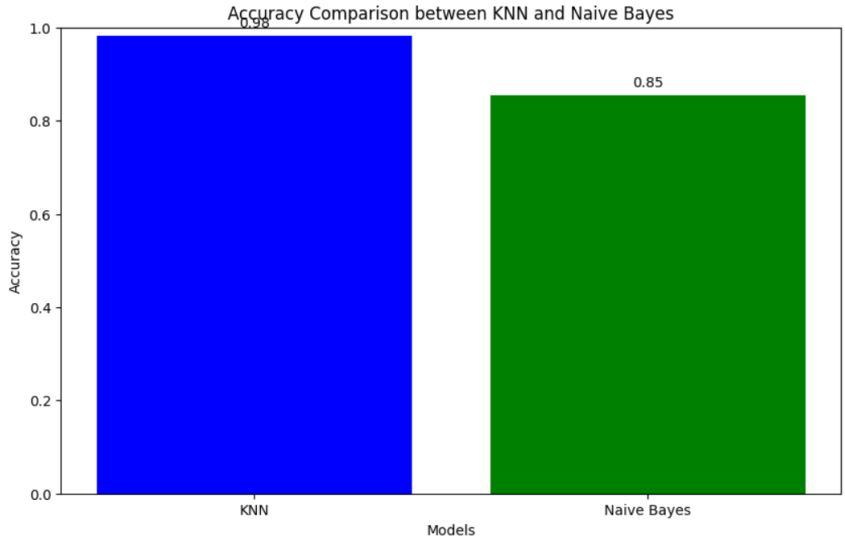


Image 1. Accuracy Comparison Between KNN and Naïve Bayes

Visual Representation

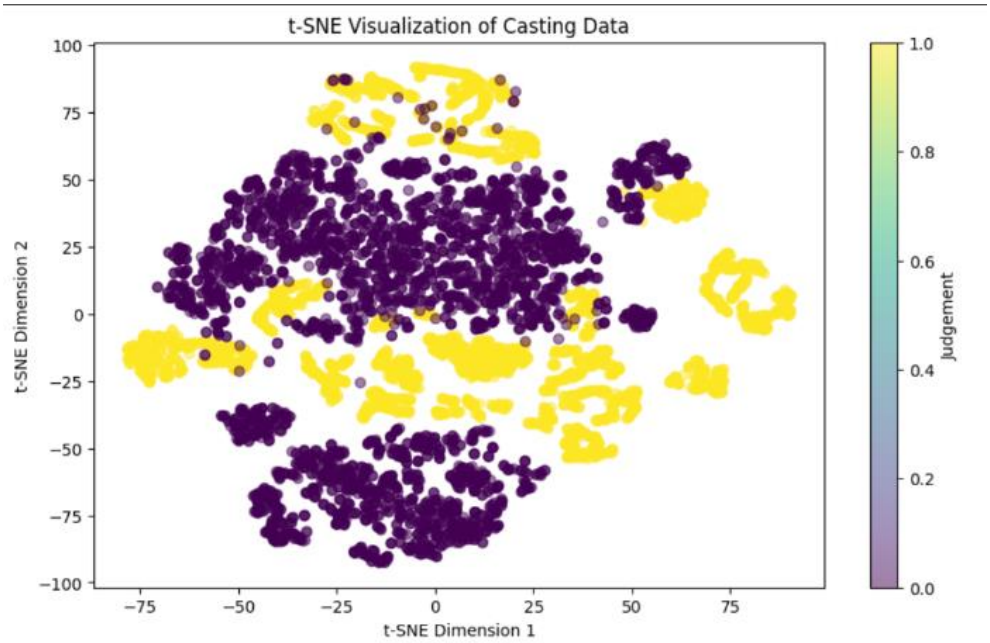


Image 2. T-SNE Visualization of Casting Data

The visualization reveals distinct clusters, indicating that the t-SNE algorithm effectively separates the defective and non-defective products based on the extracted features. This clear separation between the yellow and dark purple clusters suggests that the features used for classification are highly discriminative and capable of accurately distinguishing between defective and non-defective impeller products. Such clustering supports the reliability of the feature extraction process, including methods like HOG and LBP, which capture essential edge, gradient, and texture information. The successful application of Principal Component Analysis (PCA) to reduce dimensionality while retaining significant features further enhances the model's performance. Overall, this visualization underscores the effectiveness of the chosen pre-processing and machine learning techniques in achieving accurate defect detection.

DISCUSSION

In this research, we aimed to evaluate the effectiveness of two machine learning algorithms, K-Nearest Neighbors (KNN) and Naive Bayes, in identifying defects in impeller products. Our findings highlight the strengths and limitations of these algorithms in the context of industrial defect detection.

The KNN algorithm exhibited superior performance, achieving high accuracy, precision, recall, and F1-score. This success can be attributed to its non-parametric nature, which does not assume any underlying data distribution. KNN's capability to utilize the proximity of data points enables it to effectively classify defective and non-defective products based on feature similarity. The t-SNE visualization further supports this, showing distinct

clusters of defective (yellow) and non-defective (purple) products, indicating that the feature extraction and selection processes were successful in differentiating the two classes.

Conversely, the Naive Bayes algorithm, while still performing reasonably well, demonstrated lower accuracy, precision, and recall compared to KNN. This is likely due to its assumption of feature independence, which may not hold true in the complex data structure of casting defects. Nevertheless, Naive Bayes remains a valuable tool for defect detection due to its simplicity and speed, making it suitable for real-time applications where quick decision-making is crucial.

The t-SNE visualization revealed clear clusters corresponding to defective and non-defective products, underscoring the effectiveness of the chosen feature extraction methods such as HOG and LBP. These methods captured essential edge, gradient, and texture information, which are crucial for identifying surface defects. Principal Component Analysis (PCA) further enhanced this by reducing dimensionality, ensuring that only the most relevant features were used in model training.

The implementation of a 5-fold cross-validation technique enhanced the robustness and reliability of our models. With this method, the dataset was divided into five subsets, and the models were repeatedly trained and tested on various combinations of these subsets. This provided a comprehensive evaluation of model performance and minimized the risk of overfitting

CONCLUSIONS AND RECOMMENDATIONS

Conclusions

This study evaluated the effectiveness of K-Nearest Neighbors (KNN) and Naive Bayes algorithms in detecting defects in impeller products. The findings revealed that KNN outperforms Naive Bayes, achieving an accuracy of 98.11% compared to 85.38% for Naive Bayes. The high performance of KNN, demonstrated by superior precision, recall, and F1-score, indicates its robustness and reliability for industrial defect detection. The t-SNE visualization further supports these results, showing distinct clusters of defective and non-defective products, highlighting the effectiveness of the feature extraction and selection processes.

Recommendations

Given its high accuracy, KNN should be implemented in industrial defect detection systems to enhance quality control processes. Continuing to use and improve feature extraction and selection techniques like HOG, LBP, and PCA will ensure high-quality input data. For scenarios requiring quick decision-making, Naive Bayes remains a viable option due to its simplicity and computational efficiency; efforts should be made to improve its accuracy. Exploring hybrid models that combine KNN's accuracy and Naive Bayes' efficiency can provide a balanced defect detection solution. Additionally, ensuring that models are scalable and adaptable to various defects and products, with continuous monitoring and retraining, will maintain their effectiveness. By implementing these recommendations, industries can improve

defect detection processes, ensuring higher quality standards and operational efficiency. This research advances the use of machine learning in industrial quality control, offering practical insights and potential pathways for future advancements.

ADVANCED RESEARCH

Each study has its limitations, and this research is no exception. One of the main limitations is the dependency on the quality and diversity of the dataset. The dataset utilized in this study, despite being thorough, might not cover all possible defect scenarios in impeller products. Additionally, the algorithms' performance could vary with different types of casting defects not represented in the dataset.

Future research could address these limitations by incorporating a more diverse and extensive dataset, representing a broader range of defect types and conditions. Furthermore, Investigating the incorporation of other advanced machine learning methods, such as deep learning models, could improve the detection accuracy and generalization abilities of the system. Another area for future research is the development of real-time defect detection systems that can be seamlessly integrated into the manufacturing process. This involves optimizing the computational efficiency of the models to handle large-scale data processing and implementing robust mechanisms for continuous model learning and adaptation to new defect patterns.

Lastly, investigating the impact of different feature extraction and selection methods on model performance across various industrial applications could provide deeper insights and drive innovations in machine learning-based quality control systems.

ACKNOWLEDGMENT

I would like to thank my colleagues for their valuable suggestions and feedback during the development of this paper. Special thanks to my supervising professors for their unwavering support. Your contributions are greatly appreciated.

REFERENCES

casting product image data for quality inspection. (n.d.). Retrieved May 31, 2024, from https://www.kaggle.com/datasets/ravirajsinh45/real-life-industrial-dataset-of-casting-product

Homepage, J., Delvika, B., Nurhidayarnis, S., Rinada, P. D., Abror, N., & Hidayat, A. (2022). Perbandingan Klasifikasi Antara Naive Bayes dan K-Nearest Neighbor Terhadap Resiko Diabetes pada Ibu Hamil. *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, *2*(2), 68–75. https://doi.org/10.57152/MALCOM.V2I2.432

Penerapan Algoritma K-Means Clustering Untuk Mengetahui Kemampuan Karyawan IT | Computer Science (CO-SCIENCE). (n.d.). Retrieved May 31, 2024, from http://103.75.24.116/index.php/co-science/article/view/623

Sitepu, R., & Manohar, M. (2022). Implementasi Algoritma K-Nearest Neigbor Untuk Klasifikasi Pengajuan Kredit. *Jurnal Sistem Informasi, Teknik Informatika Dan Teknologi Pendidikan*, *1*(2), 49–56. https://doi.org/10.55338/JUSTIKPEN.V1I2.6

View of Analisis Sentimen pada Media Sosial dengan Teknik Kecerdasan Buatan Naïve Bayes: Kajian Literatur Review. (n.d.). Retrieved May 31, 2024, from https://journal.mediapublikasi.id/index.php/oktal/article/view/2944/1371

View of Analisis Sentimen Relokasi Ibukota Nusantara Menggunakan Algoritma Naïve Bayes dan KNN. (n.d.). Retrieved May 31, 2024, from https://jkomtekinfo.org/ojs/index.php/komtekinfo/article/view/330/160

View of Klasifikasi Bidang Minat Mahasiswa Elektronika Dalam Menentukan Topik Tugas Akhir Menggunakan Algoritma Naïve Bayes Classifier (Studi Kasus: Prodi Pendidikan Teknik Informatika FT-UNP). (n.d.). Retrieved May 31, 2024, from https://www.jptam.org/index.php/jptam/article/view/6813/5671

View of KLASIFIKASI DIAGNOSA PENYAKIT DIABETES DENGAN METODE NAIVE BAYES BERBASIS WEB. (n.d.). Retrieved May 31, 2024, from https://ojs.ninetyjournal.com/index.php/JKBTI/article/view/35/26

View of METODE NAIVE BAYES UNTUK KLASIFIKASI MASA STUDI SARJANA. (n.d.). Retrieved May 31, 2024, from http://teknologipintar.org/index.php/teknologipintar/article/view/385/370

View of Pemanfaatan Kecerdasan Buatan Dalam Mendukung Pembelajaran Mandiri. (n.d.). Retrieved May 31, 2024, from https://ejournal.uika-bogor.ac.id/index.php/EDUCATE/article/view/14843/4618