Building of Informatics, Technology and Science (BITS)
Volume 6, No 4, March 2025 Page: 2193−2203
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i4.6510

Classification of Key and Time Signature in Western Musical Notation by using CRNN Algorithm with Bounding Box

Dennis Adiwinata Irwan Soeroso*, Sri Winarno, Ardytha Luthfiarta, Firda Ayu Dwi Aryanti

Faculty of Computer Science, Informatics Engineering, Dian Nuswantoro University, Semarang, Indonesia
Email: [1,*]111202113225@mhs.dinus.ac.id, [2]sri.winarno@dsn.dinus.ac.id, [3]ardytha.luthfiarta@dsn.dinus.ac.id, [4]firdaaryanti65@gmail.com
Correspondence Author Email: 111202113225@mhs.dinus.ac.id
Submitted: **24/12/2024**; Accepted: **26/12/2025**; Published**: 01/03/2025**

Abstract-This research seeks to employ the Convolutional Recurrent Neural Network (CRNN) algorithm to develop a method for classifying key and time signatures from sheet music images. The research design involved compiling a dataset of 285 sheet music images, which includes 15 types of key signatures and 19 types of time signatures. The methodology encompasses annotation using the bounding box technique, image preprocessing, and applying the CRNN model for classification using K-Fold Cross Validation because of the limited dataset. Then, the model is evaluated using the Multi Class Confusion Matrix and performance metrics. The primary findings of this study reveal that the developed model achieves 96% accuracy in key signature classification and 95% in time signature classification when utilizing bounding boxes. Conversely, the absence of bounding boxes substantially negatively impacted the accuracy of key signature classification, resulting in only a 58% accuracy rate. Time signature classification performed even worse, with an accuracy of just 19%. This research highlights the substantial accuracy enhancements achievable by incorporating bounding boxes. Therefore, we anticipate that this research will help singers, especially those in choirs, to understand and express music better using existing technologies while enhancing the accuracy of optical music recognition using the CRNN model.

Keywords: Bounding Box; CRNN Algorithm; Key Signature; Optical Music Recognition; Time Signature

1. INTRODUCTION

A choir consists of a group of singers, including female voices categorized as soprano, mezzo-soprano, and alto, and male voices categorized as tenor, baritone, and bass [1], [2]. A conductor usually leads the choir itself. Songs that are typically performed in the choir can be church songs, regional songs, or modern songs [2]. However, these songs are generally written in Western musical notation. This can be understood because Western musical notation is the most popular type of notation in music. A musician who is mainly in the classical field, be it musical instruments or voice, must feel more familiar with this type of notation [3]. Numbered musical notation is easier to learn and write than Western musical notation, used from elementary to high school in Indonesia [4]. Because of its simplicity, we are often more familiar with numbered musical notation in our music education. This can be understood partly because the system of writing music scores that are introduced and widely used, for example, in Sundanese gamelan, resembles numbered musical notation more than Western musical notation [5].

In the world of choirs, where singers come from many different backgrounds, it is often difficult to understand the music score in the form of Western musical notation, even to determine the key and time signature, which are essential for us to understand the song better [2]. Because of this limitation, we need to utilize existing technology to help classical singers, including choirs in Indonesia, to more easily understand how a song can be sung with a notation that is easier to understand and more suitable for their musical background [6].

In the application of technology to translate music notation in the form of Western musical notation, we can use Optical Music Recognition (OMR) [7]. OMR research seeks to find how computers can recognize musical scores effectively. Various studies have been conducted to improve the accuracy and efficiency of OMR, including developing a new dataset that includes examples of music theory from the 19th century [8] and using deep neural networks for composer identification [9]. Other studies have also pointed out the challenges in synthesizing training data for OMR, which can affect recognition results [10]. In addition, end-to-end approaches in OMR have been introduced to improve the recognition of monophonic music documents [9] and the development of graph models for more efficient OMR [11].

Key signature (chord) and time signature (tempo) are key components in music that greatly influence the interpretation and listening experience. Chords provide the underlying harmonic structure of the melody [12]. At the same time, tempo determines the speed and rhythm of the song, which is crucial for conveying the emotion and character of a piece of music. Research shows that a good understanding of chords and tempo can enhance the user experience in music applications and assist singers in expressing music better [13]. Therefore, accurately classifying these two elements is crucial in OMR and other music applications.

In the context of OMR, the Convolutional Recurrent Neural Network (CRNN) algorithm is one of the commonly used algorithms. CRNN itself is a combination of the Convolutional Neural Network (CNN) and the Recurrent Neural Network (RNN) [14]. CNN is often used to perform symbol analysis because this algorithm is suitable for extracting spatial characteristics from images of musical scores [4]. At the same time, RNN is used to handle the sequential nature of music, where this algorithm is good to use on data that predicts the sequence of musical notes [15].

Building of Informatics, Technology and Science (BITS)
Volume 6, No 4, March 2025 Page: 2193−2203
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i4.6510

However, although several OMR studies have been conducted to recognize music notation, there is still a gap in the automatic classification of key and time signatures from sheet music images, especially in solving the problem with the choir in Indonesia. Previous research has focused chiefly on music notation recognition without paying particular attention to the classification of key and time signature, which are essential elements in music interpretation, such as research by Rios-Vila et all; that using CRNN with KERN and BEKERN encoding without paying attention to the semantic music grammar and not show the pitch and tempo notation that much familiar used in Indonesia [16]. Previous research by Edirisooriya, Sachinda et all. Has also been conducted on end-to-end polyphonic OMR using the CRNN algorithm, resulting in two models, FlagDecoder and RNNDecoder, resulting in a relatively good accuracy but cannot solve the problem that we try to solve in this research that is to help identify the key and time signature on western musical notation to numbered music notation [17]. Another example is research by M. Alfaro-Contreras et all. They have demonstrated progress in music notation recognition but have yet to explicitly address the challenge of identifying key and time signatures from complex scores [9]. Although OMR has progressed in the research by A. Ríos-Vila et all, there are still challenges in recognizing more subtle musical elements, such as key and time signatures, simply because the datasets that are used are large and much more complex, so the computational time also impacted the process [18]. In other research, P. Torras et al. used a handwritten notation that was not commonly used to write musical notations in Indonesia. This complex algorithm is also unsuitable for a much smaller dataset being used [19].

This research utilizes the CRNN algorithm to identify key and time signature sheet music images. It also performs the annotation process using bounding boxes, which is inspired by the approach used in research on object detection in CT-scan images, where bounding boxes are used to mark relevant areas in the image to improve detection accuracy [10]. Utilizing more advanced OMR technology is expected to overcome the existing limitations and provide a more effective solution for choir singers, especially those without a musical background. This research aims to develop a better method of key and time signature classification to assist singers in understanding and expressing music better.

2. RESEARCH METHODOLOGY

2.1 Research Stages

The main objective of this OMR research is to classify a musical notation's key and time signature into a numbered musical notation that is much more understandable in Indonesia. Several steps are needed to support this objective, as shown in Figure 1 below.

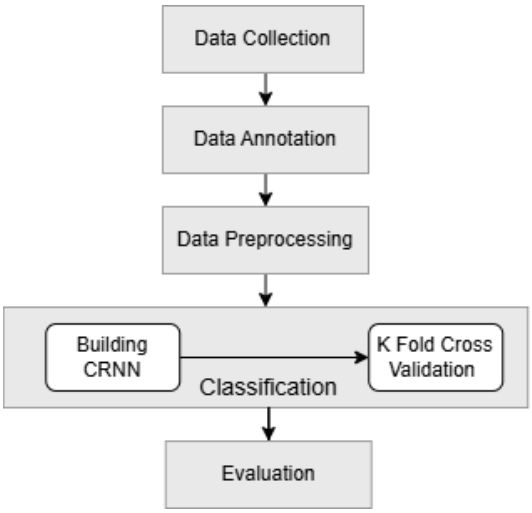

Figure 1. Research Flow

As shown in Figure 1, the first step is to collect the data and then annotate it into several classes. The third step is to preprocess the data to enrich the dataset. The fourth step is the classification, which was processed to train the data using CRNN and K-fold cross-validation. Lastly, the fifth step is to evaluate the data using a multi class confusion matrix and performance metric.

2.2 Data Collection

In this research, the first process is to collect images of block note scores that only contain the pitch and tempo. The data was collected manually with the help of tools for writing Western musical notation, namely the Musescore4 tool, which was chosen because it was a free tool that many musicians use [20]. Following this study's objectives, the data comes from 15 types of key signatures and 19 types of time signatures commonly used in the world of music, for a total of 285 images in PNG format.

Building of Informatics, Technology and Science (BITS)
Volume 6, No 4, March 2025 Page: 2193−2203
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i4.6510

2.3 Data Annotation

The next step in this research is to annotate the data. The data annotation process in this research was carried out using the LabelImg tool to mark objects with the bounding box method, which aims to prepare a dataset consisting of 15 key signature classes and 19-time signature classes. The annotated dataset includes representative images, where the research team manually tagged each image to ensure accuracy and consistency. After the annotation process, validation is performed to ensure the quality and accuracy of the tagging so that the resulting dataset can be effectively used in training machine learning models for key and time signature recognition.

2.4 Data Preprocessing

After the data is collected and annotated, the next step is preprocessing the image data. This process starts with image reading using OpenCV, which converts the image into grayscale format. This conversion is intended to simplify the data, enabling the model to concentrate on critical features without being influenced by color details [21]. Subsequently, the images obtained from the dataset were resized to 32x32 pixels. This size was chosen to ensure consistency in the model input, which is essential for efficient training. The smaller size also helps to reduce the computation time and memory required during training [22].

The bounding boxes designated during the annotation process isolate regions of interest (ROIs) for both the key and time signature, a procedure commonly referred to as segmentation. Each ROI is subsequently stored in a distinct array for model training. This step is crucial to ensure that the model focuses on relevant features while disregarding extraneous information [23]. To enhance dataset diversity and mitigate the risk of overfitting, data augmentation techniques are employed, including rotation, flipping, and rescaling of the images [24]. This augmentation seeks to enhance the dataset by introducing more significant variability, thereby enabling the model to recognize patterns across diverse conditions. Additionally, it aims to mitigate the risk of overfitting, particularly in scenarios characterized by limited dataset size [25].

2.5 Classification

In the classification stage, the CRNN algorithm is used to identify the key and time signature. This model was chosen for its ability to capture the spatial features of the sheet music image through a convolution layer, followed by a recurrent layer that can capture temporal information [15]. This approach is particularly suitable for pattern recognition in data with a two-dimensional structure, such as images, where temporal information also plays an essential role in a musical context.

The use of CRNN in OMR is based on its ability to integrate spatial and temporal information effectively. CRNNs can improve the accuracy of music notation recognition by utilizing the hierarchical structure of music data, which includes the relationships between different musical elements in a score [26]. In addition, CRNNs are also capable of handling variations in the visual representation of music notation, such as differences in size, orientation, and lighting, which are often challenges in automatic music notation recognition [15].

The CRNN model, the subject of this study, comprises multiple layers, including two convolution layers (Conv2D) and a pooling layer (MaxPooling2D) that reduces the dimensionality of the resulting features. Each convolution layer applies a ReLU (Rectified Linear Unit) activation function to introduce nonlinear elements into the model. Following the convolution and pooling stages, the model proceeds with a batch normalization layer designed to enhance stability and efficiency during training. The model culminates with multiple dense layers and a dropout layer in between them to prevent any overfitting in the model [27]The final layer employs a softmax activation function to generate classification probabilities for each key and time signature. After the algorithm has been built, the next step is to use K Fold cross validation, as shown in Figure 2 below.

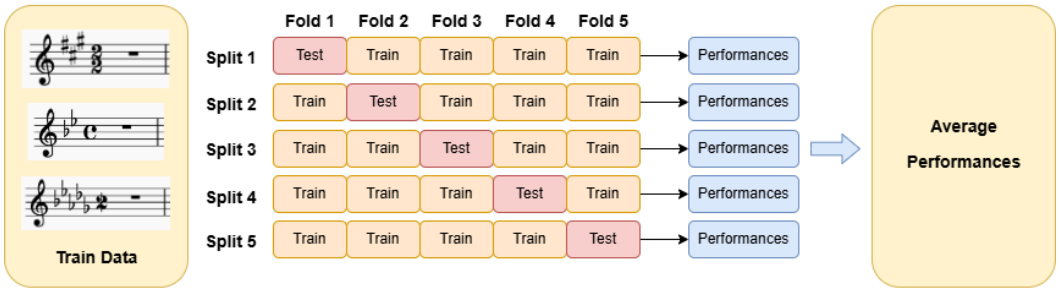

Figure 2. K Fold Cross Validation Diagram

Given that the dataset is not very large, K Fold cross validation is applied to ensure that each image is used to train and test the model [28]. This model is chosen because K Fold Cross Validation is suitable for relatively small datasets [29]. As in Figure 2, the K value in this study was set at 5, which means the dataset was divided into five subsets. The research team conducts a fivefold training protocol for the model, wherein four distinct subsets of data are utilized for training purposes while reserving one subset for the validation and testing phase. This methodological approach ensures that each subgroup maintains an equivalent number of cases, thereby enhancing the robustness and

Building of Informatics, Technology and Science (BITS)
Volume 6, No 4, March 2025 Page: 2193−2203
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i4.6510

reliability of the evaluation process. In this way, each image in the dataset has an equal opportunity to contribute to the training and evaluation of the model [30]. In the final stage, the results of the five trials are averaged to obtain a more accurate and robust model, which helps get a more precise estimation of the model's performance and reduces the variability that may occur due to unequal data sharing [31].

2.6 Evaluation

Following the model's training, this research's final stage is evaluation. This evaluation measures the model's performance in classifying each loop's key and time signatures using several methods, such as the confusion matrix and the performance metric (classification report). The confusion matrix itself is a method applied to find and compare information from actual results with classification prediction results [32]. The confusion matrix is used to identify the anomaly in the predicted data of the model used in this research so that we can identify the main reason behind that and compare. This research uses Multi Class Confusion Matrix as in Table 1, meaning that each instance is only labeled as one class [33].

Table 1. Multi Class Confusion Matrix for *i* Class

	Predicted Class 1	Predicted Class 2	⋯	Predicted Class *i*
Actual Class 1	TP_1	$FP_{1,2}$	⋯	$FP_{1,i}$
Actual Class 2	$FN_{2,1}$	TP_2	⋯	$FP_{2,i}$
⋮	⋮	⋮	⋱	⋮
Actual Class *i*	$FN_{i,1}$	$FN_{i,2}$	⋯	TP_i

As shown in the table 1 confusion matrix has four main categories: True Positive (TP) is the count of correctly predicted instances for class C_i, True Negative (TN) is the sum of all the cases that do not belong to class C_i and were predicted as not belonging to class C_i, False Positive (FP) is the sum of instances that belong to class C_i but were incorrectly predicted as class C_i, and lastly False Negative (FN) is the sum of cases that belong to class C_i but were predicted as any other class C_j. The formula used to calculate these four values for each class (C) is shown in equations number 1 to number 4. In these four equations, i is denoted as the index of the class for which the matrix will be calculated, and then the value j is used to represent an index other than the C_i class. Another similar term is k, which is used like j but to count all classes other than the C_i class when calculating FP. The other existing terms are m and n, which are used to calculate TN, but they represent rows and columns other than the C_i class, respectively [33].

$$TP_{C_i} = Confusion\ Matrix[C_i, C_i] \tag{1}$$

$$FN_{C_i} = \sum_{j \neq i} Confusion\ Matrix[C_i, C_j] \tag{2}$$

$$FP_{C_i} = \sum_{k \neq i} Confusion\ Matrix[k, C_i] \tag{3}$$

$$TN_{C_i} = \sum_{m \neq i} \sum_{n \neq i} Confusion\ Matrix[m, n] \tag{4}$$

After identifying the confusion matrix, we will calculate commonly used performance metrics such as accuracy, which refers to the proportion of correctly classified data relative to the total data. In contrast, precision measures the ratio of accurate positive predictions to the number of optimistic predictions [34]. Recall evaluates the ratio of correctly identified positive samples compared with the total number of positive samples. The F1 score is a metric that combines precision and recall, offering a singular measure of classification efficacy [35]. The four matrices can be calculated based on the formula in number 5 until number 8.

$$accuracy = \frac{TN+TP}{TN+FP+TP+FN} \tag{5}$$

$$precision = \frac{TP}{TP+FP} \tag{6}$$

$$recall = \frac{TP}{TP+FN} \tag{7}$$

$$F1\ Score = (2 * precision * recall)/(precision + recall) \tag{8}$$

3. RESULT AND DISCUSSION

3.1 Data Collection

At this stage, data collection of Western musical notation images containing key and time signatures is carried out, which will be used in this study. Images are obtained by using the musescore4 tool to create notations, which are then screenshots to be saved into PNG format. The data comes from 15 types of key signatures and 19 types of time signatures commonly used in the world of music, so the total dataset used is 285 images. Figure 3 below is an example of the data used in this research.

Building of Informatics, Technology and Science (BITS)
Volume 6, No 4, March 2025 Page: 2193−2203
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i4.6510

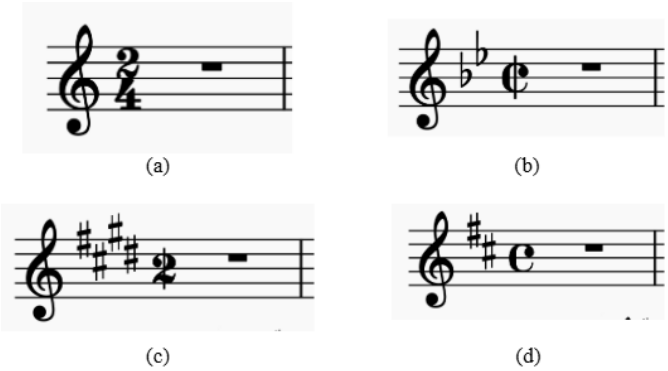

Figure 3. Picture Dataset Example of Key and Time Signature Notation

Figure 3 is a dataset example of Western musical notation that contains key and time signatures. Figure 3 (a) is a dataset with a C key signature and a 2/4 time signature. Figure 3 (b) is a dataset with a B flat (Bes) key signature and an alla breve or 2/2 time signature. Next, Figure 3 (b) is a dataset with an E key signature and a cut time or 2/2 time signature. Lastly, Figure 3 (b) is a dataset with a D key signature and a common or 4/4 time signature.

3.2 Data Annotation

The data annotation process in this study was conducted using the LabelImg tool, which was used to apply the bounding box technique. The data was manually tagged to ensure accuracy and consistency. The key signature classes include variations such as A, A_flat, B, B_flat, C, C_flat, C_sharp, D, D_flat, E, E_flat, F, F_sharp, G, and G_flat,. The tempo class includes categories such as 2-2, 2-4, 3-2, 3-4, 3-8, 4-2, 4-4, 4-8, 5-4, 5-8, 6-4, 6-8, 7-8, 9-8, 12-8, alla_breve, common, cut_time, and cut_triple. Figure 4 is an example of this annotation process of the C-sharp key signature class and 6-4 time signature class.

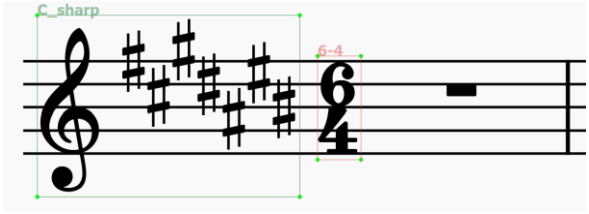

Figure 4. Annotated Example of Key and Time Signature Notation

3.3 Data Preprocessing

The preprocessing process in this research aims to prepare the music image data to be more structured and ready for use in the machine learning model. Data in the form of music score images are taken from a directory that has been provided. This directory consists of two main folders: Picture for images and Annotation for annotations in XML format. These annotations contain information about each image's key and time signature's position (bounding box). The main objective of this process is to generate ROIs of every image.

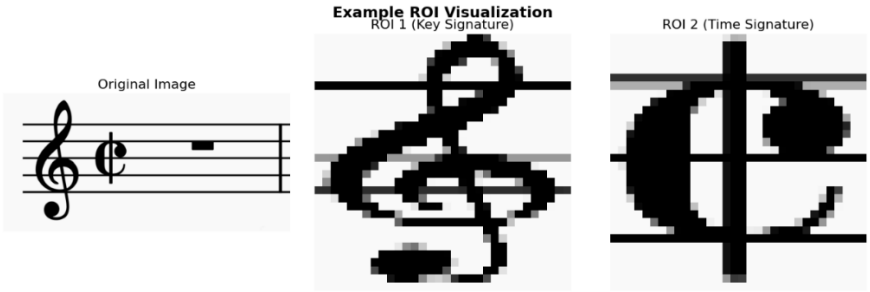

Figure 5. Annotated Example of Key and Time Signature Notation

Figure 5 above shows the process for extracting the ROI for the image dataset with a C key signature and an alla breve time signature. The bounding box specified in the XML file is employed to crop the pertinent portion of the image. Two ROIs are extracted for each image: one for the key signature and one for the time signature. If the number of bounding boxes is fewer than two, the data is bypassed to maintain consistency. After extraction, the extracted data will be normalized and resized to 32x32 pixels to maintain consistency.

Building of Informatics, Technology and Science (BITS)
Volume 6, No 4, March 2025 Page: 2193−2203
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i4.6510

3.4 Result

The evaluation results of the key and time signature classification model are shown in the performance metrics table and confusion matrix presented in the figure. The model achieved an overall accuracy of 96% for the key signature classification in Figure 6. The evaluation metrics used include precision, recall, and F1-score, which provide a comprehensive overview of the model's performance in identifying relevant classes.

	precision	recall	f1-score	support
A	1.00	0.80	0.89	5
A_flat	1.00	1.00	1.00	5
B	1.00	1.00	1.00	2
B_flat	1.00	1.00	1.00	3
C	1.00	1.00	1.00	6
C_flat	1.00	1.00	1.00	3
C_sharp	1.00	1.00	1.00	4
D	1.00	1.00	1.00	3
D_flat	1.00	0.50	0.67	2
E	0.67	1.00	0.80	2
E_flat	1.00	1.00	1.00	6
F	1.00	1.00	1.00	5
F_sharp	0.86	1.00	0.92	6
G	1.00	1.00	1.00	2
G_flat	1.00	1.00	1.00	3
accuracy			0.96	57
macro avg	0.97	0.95	0.95	57
weighted avg	0.97	0.96	0.96	57

Figure 6. Classification Report of Key Signature with Bounding Box

The key signature classification in Figure 6 showed excellent results, with the model achieving an overall accuracy of 96%. The evaluation results table shows that most of the classes have high precision and recall values, where class A has a precision of 1.00, recall of 0.80, and F1-score of 0.89, indicating that although the model can identify all positive instances well, there are 20% of class A instances that are not detected. Classes A_flat, B, and C_sharp achieved perfect scores with a precision and recall of 1.00, indicating the model's excellent ability to recognize these classes. However, classes D_flat and F_sharp show more mixed results, with D_flat having a precision of 1.00 and recall of 0.67 and F_sharp with a precision of 0.86 and recall of 1.00, indicating that although the model can recognize most instances, there are still some that go undetected. On the other hand, Class E has a lower F1 score of 0.67, indicating challenges in identification. This shows a substantial difference from the model that does not use a bounding box.

Classification Report for Pitch:

	precision	recall	f1-score	support
A	0.86	0.46	0.60	329
A_flat	0.85	0.46	0.59	330
B	0.92	0.45	0.60	329
B_flat	0.90	0.48	0.62	329
C	0.13	0.99	0.23	330
C_flat	0.93	0.45	0.61	329
C_sharp	0.89	0.47	0.61	329
D	0.84	0.39	0.53	330
D_flat	0.86	0.47	0.61	329
E	0.82	0.47	0.60	329
E_flat	0.95	0.46	0.62	330
F	0.91	0.43	0.59	329
F_sharp	0.92	0.51	0.66	329
G	0.88	0.51	0.64	330
G_flat	0.88	0.49	0.63	329
accuracy			0.50	4940
macro avg	0.84	0.50	0.58	4940
weighted avg	0.84	0.50	0.58	4940

Figure 7. Classification Report of Key Signature without Bounding Box

The key and time signature classification model evaluation results show significant differences between using and without bounding boxes, as shown in Figure 7. In the key signature classification, the model using bounding boxes achieved 96% accuracy, with high precision and recall values for most classes. Class A, for example, has a precision of 1.00 and recall of 0.80, indicating the model's excellent ability to recognize this class. However, in the latest results without bounding boxes, the overall accuracy for key signature classification only reached 58%, with lower precision and recall values for many classes. Class A, for example, has a precision of 0.86 and a recall of 0.46, indicating that the model struggles to identify positive instances consistently. Next, a multi-class confusion matrix, as shown in Figure 8 below, is generated to identify the data. The resulting confusion matrix will show a significant difference between models that use bounding boxes and models that do not.

Building of Informatics, Technology and Science (BITS)
Volume 6, No 4, March 2025 Page: 2193−2203
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i4.6510

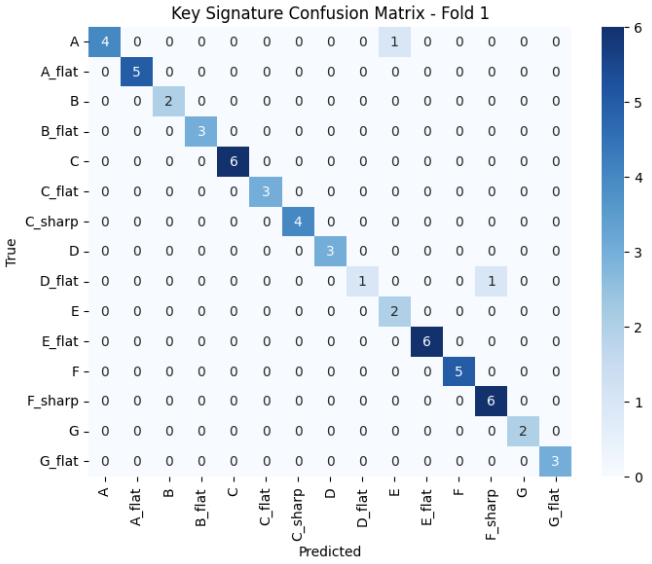

Figure 8 Confusion Matrix of Key Signature with Bounding Box

The multi-class confusion matrix for key signature classification in Figure 8 shows that most predictions are correct, although there are misclassifications, especially between classes A and D flat. The model predicts class A to be class E_flat, which is understandable because the key signatures of A and E_flat have similarities: the flat and sharp (chromatic accidental) signs are both three. As seen in Figure 8, the model also failed to predict class D_flat, which is misclassified as F_sharp. To simplify our understanding, in Figure 3, where class B_flat and D have the same number of chromatic accidentals, both have two flat or sharp symbols. Similarities like this cause some errors in the model because the data has symbols that are very similar to each other.

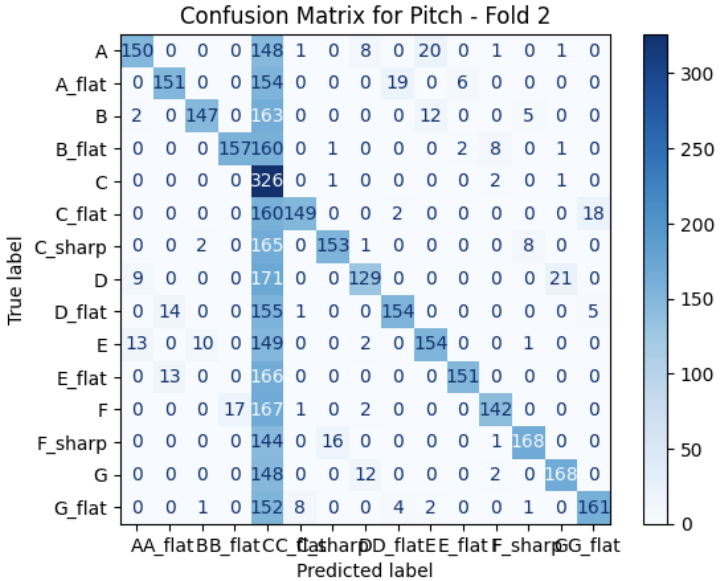

Figure 9. Confusion Matrix of Key Signature without Bounding Box

Figure 9 shows that all classes are misclassified with class C. This is the main challenge of models that do not use bounding boxes. The problem is mainly because the C key signature only uses one G-clef symbol. Because this is the basis of all key signatures, every image must have a G-clef symbol, so the model misclassifies a lot.

Regarding time signature classification, the bounding box model achieved 95% accuracy, with classes 3-4 having a precision of 0.86 and a recall of 1.00.

Next, the tempo classification showed promising results, as shown in Figure 10, with the model achieving an overall accuracy of 95%. In contrast, a model that does not use a bounding box shows many misclassified data.

Building of Informatics, Technology and Science (BITS)
Volume 6, No 4, March 2025 Page: 2193−2203
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i4.6510

	precision	recall	f1-score	support
12-8	1.00	1.00	1.00	2
2-2	1.00	1.00	1.00	3
2-4	1.00	1.00	1.00	1
3-2	1.00	1.00	1.00	2
3-4	0.86	1.00	0.92	6
3-8	1.00	0.75	0.86	4
4-2	1.00	1.00	1.00	2
4-4	1.00	1.00	1.00	4
4-8	1.00	1.00	1.00	2
5-4	1.00	0.80	0.89	5
5-8	1.00	1.00	1.00	2
6-4	1.00	1.00	1.00	3
6-8	1.00	1.00	1.00	2
7-8	0.83	1.00	0.91	5
9-8	0.67	0.67	0.67	3
alla_breve	1.00	1.00	1.00	2
common	1.00	1.00	1.00	2
cut_time	1.00	1.00	1.00	5
cut_triple	1.00	1.00	1.00	2
accuracy			0.95	57
macro avg	0.97	0.96	0.96	57
weighted avg	0.95	0.95	0.95	57

Figure 10. Classification Report of Time Signature with Bounding Box

As seen in Figure 10, the evaluation results table shows that most classes have high precision and recall values, but some still need to achieve perfect scores. Class 3-4 had a precision of 0.86 and a recall of 1.00, resulting in an F1-score of 0.92. Although the model can correctly identify all positive instances, some instances are not detected, which may be due to variations in the visual representation of the tempo notation. Classes 7-8 showed lower results, with a precision of 0.83 and recall of 0.91, indicating that the model needed help recognizing these classes consistently. Classes 9-8 had similar precision and recall values of 0.67, indicating that the model could not distinguish this class well from the others.

Classification Report for Tempo:

	precision	recall	f1-score	support
3/4	0.15	0.03	0.06	261
6/4	0.26	0.08	0.13	248
5/8	0.07	0.02	0.03	266
3/2	0.15	0.05	0.08	254
6/8	0.16	0.10	0.13	249
4/8	0.13	0.05	0.07	235
7/8	0.14	0.06	0.08	275
2/4	0.26	0.09	0.14	255
9/8	0.14	0.10	0.12	530
5/4	0.21	0.07	0.10	265
3/8	0.24	0.06	0.10	245
4/2	0.21	0.07	0.11	275
12/8	0.66	0.27	0.38	286
4/4	0.27	0.12	0.17	534
2/2	0.18	0.75	0.29	762
accuracy			0.19	4940
macro avg	0.21	0.13	0.13	4940
weighted avg	0.21	0.19	0.15	4940

Figure 11. Classification Report of Time Signature without Bounding Box

However, in the latest results without bounding boxes in Figure 11, the accuracy for time signature classification only reaches 19%, with classes 3/4 having a precision of 0.15 and recall of 0.03. This shows that the model needs to recognize time signature classes with bounding boxes, possibly due to the visual similarity and variation in time signature notation. The difference between Figure 10 and Figure 11 is in the Figure 11 classes, such as alla breve, common, cut_time, and cut_triple, which are often depicted the same as 4/4 or 2/2 time signatures, which is why the model has much support for these two classes.

Building of Informatics, Technology and Science (BITS)
Volume 6, No 4, March 2025 Page: 2193−2203
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i4.6510

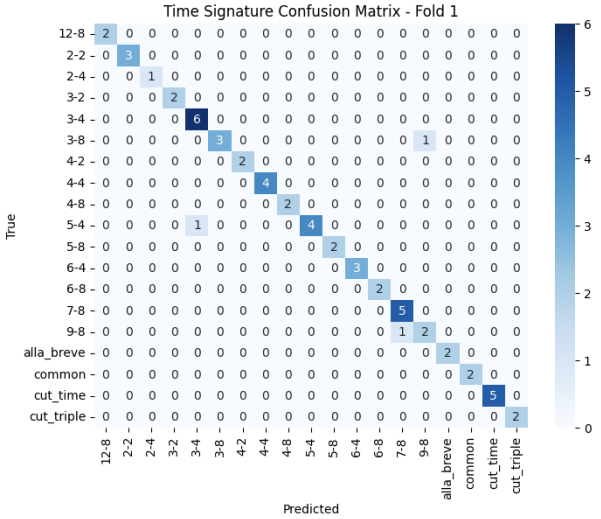

Figure 12. Confusion Matrix of Time Signature with Bounding Box

The confusion matrix for time signature classification in Figure 12 shows a similar pattern to the key signature's confusion matrix. Most predictions are correct, but there are still some misclassifications. For example, classes 3-8, 5-4, and 9-8 show significant misclassification, where the model misidentifies some instances as other classes. This indicates the challenge in distinguishing between these classes because of the shape similarity for each class.

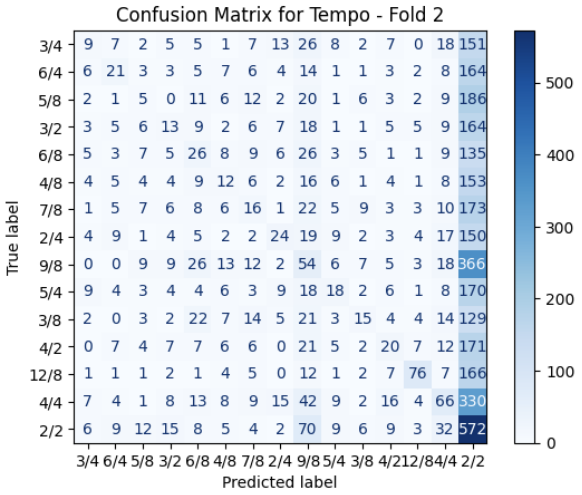

Figure 13. Confusion Matrix of Time Signature without Bounding Box

The confusion matrix for time signature classification with bounding boxes in Figure 12 shows that although most predictions are correct, some misclassifications must be noted. In contrast, the confusion matrix for time signature classification without bounding boxes in Figure 13 shows a worse pattern, where the model often misclassified the classes, indicating the challenge in distinguishing between these classes. The confusion matrix in Figure 13 is mainly classified data as 2/2.

3.5 Discussion

Based on the results of this study, it is found that the best accuracy for key signature calculation was 96%, while for time signature, it was 95%. Another similar study conducted on piano sheet music images using the CRNN model with KERN, KERN-SP, and BEKERN encodings resulted in accuracies of 92.7%, 90.8%, and 93.9%, respectively [16]. Compared to this study, the use of CRNN with a bounding box shows a significant increase in accuracy, which is about 2-3%, indicating that the model applied in this study is very suitable for symbol recognition in Western musical scores. In addition, previous research has also been conducted on end-to-end polyphonic OMR using the CRNN algorithm, which resulted in two models: FlagDecoder and RNNDecoder. The study used evaluation with SER, where the accuracy of CRNN with FlagDecoder reached 93.33% for time signature and 90.18% for key signature. Meanwhile, CRNN with RNNDecoder obtained results of 96.08% for time signature and 94.36% for key signature [17]. From these results, it can be concluded that using CRNN with bounding box annotations is very suitable, especially for key signatures, because it can increase accuracy compared to the previous research. As for time signature, there is a difference in accuracy of about 1%. Other research by M. Alfaro-Contreras that used the

Building of Informatics, Technology and Science (BITS)
Volume 6, No 4, March 2025 Page: 2193−2203
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i4.6510

CRNN algorithm found that the result of using base CRNN without other modeling is around 92.35% and 94.6% for CRNN that has been applied using the OMR-specific method [9]. These results show that using bounding boxes in model training significantly improves accuracy and classification ability, especially in the context of key and time signature recognition in Western musical notation. This emphasizes the importance of appropriate figure-processing techniques in improving model performance in complex classification tasks.

4. CONCLUSION

Based on this research, it can be concluded that using the CRNN algorithm in classifying key and time signature from musical score images shows promising results. The developed model achieved a high accuracy of 96% for key signature classification and 95% for time signature classification when using the bounding box annotation method. These results show that an approach that integrates symbol prediction can improve the effectiveness of music notation recognition. The model achieved a key signature accuracy of 96% and a time signature accuracy of 95%, which is a notable improvement compared to other studies. For instance, previous research on piano sheet music using CRNN with various encodings reported accuracies ranging from 90.8% to 93.9%. Additionally, end-to-end polyphonic OMR studies using CRNN models like FlagDecoder and RNNDecoder achieved accuracies of 90.18% to 96.08% for key and time signatures. The 2-3% increase in accuracy observed in this study underscores the effectiveness of using bounding box annotations, particularly for key signatures, while showing a slight improvement for time signatures. However, this study also identified variations in performance between classes, where some classes, such as A and E_flat, performed better than others, such as E. This indicates that the visual complexity of musical notation can affect the model's classification ability. In addition, the comparison between the results with and without the use of bounding boxes shows that the correct annotation technique dramatically affects the model's accuracy. Without bounding boxes, the accuracy for key signature classification drops significantly, reaching only 58%, and for time signature classification, the accuracy drops to 19%. This research emphasizes the importance of developing better methods in music notation recognition, especially in the context of OMR, to help choir singers, especially those from different backgrounds. Utilizing existing technologies is expected to overcome the limitations in understanding music notation and improve the user experience in interacting with music. Further research can be done to explore more sophisticated data augmentation techniques, develop more robust models, and apply this method to more extensive and diverse datasets. This is expected to improve accuracy and efficiency in music notation recognition and significantly contribute to the development of OMR technology.

ACKNOWLEDGMENT

I would like to express my deepest gratitude to God Almighty and my advisor, family, and college friends for their guidance and support while writing this article. I would also like to thank my friends in the Gita Dian Nuswa choir at Dian Nuswantoro University for their assistance in annotating the image and for teaching me how to distinguish between the various types of key signatures and read Western musical notation. Moving forward, I aspire to develop this research further to create a translator that converts Western musical notation to numbered musical notation, which Indonesians and most choir members at Dian Nuswantoro University generally more widely understand.

REFERENCES

[1] K. B. Pratama, S. Suyanto, and E. Rachmawati, "Human Vocal Type Classification using MFCC and Convolutional Neural Network," in 2021 International Conference on Communication & Information Technology (ICICT), IEEE, Jun. 2021, pp. 43–48. doi: 10.1109/ICICT52195.2021.9568474.

[2] J. K. L. Dimpudus, A. M. Sambul, and A. S. M. Lumenta, "Transliteration Block Notation Application Into Number Notation Using The MusicXML Format," Jurnal Teknik Informatika, vol. 7, no. 1, pp. 75–82, Jan. 2022, doi: https://doi.org/10.35793/jti.v17i1.36298.

[3] R. Broude and M. Cyr, "The Emergence of Efficient Musical Texts during the Age of Reason," Textual Cultures, vol. 15, pp. 159–94, 2022, doi: 10.14434/tc.v15i1.35540.

[4] Q. Wang, L. Zhou, and X. Chen, "Kernel Density Estimation and Convolutional Neural Networks for the Recognition of Multi-Font Numbered Musical Notation," Electronics (Switzerland), vol. 11, no. 21, Nov. 2022, doi: 10.3390/electronics11213592.

[5] M. Sasaki and J. Masunah, "A Review of The Sundanese Scale Theory," Harmonia: Journal of Arts Research and Education, vol. 21, no. 2, pp. 318–329, Dec. 2021, doi: 10.15294/harmonia.v21i2.32995.

[6] N. Li, "Generative Adversarial Network for Musical Notation Recognition during Music Teaching," Comput Intell Neurosci, vol. 2022, 2022, doi: 10.1155/2022/8724688.

[7] J. Calvo-Zaragoza, J. Hajic, and A. Pacha, "Understanding Optical Music Recognition," ACM Comput Surv, vol. 53, no. 4, Sep. 2020, doi: 10.1145/3397499.

[8] Fabian C. Moss, Maik K¨oster, N´estor N´apoles L´opez, and David Rizo, "Proceedings of the 4th International Workshop on Reading Music Systems," Challenging sources: a new dataset for OMR of diverse 19th-century music theory examples , Nov. 2022, doi: 10.48550/arXiv.2211.13285.

Building of Informatics, Technology and Science (BITS)
Volume 6, No 4, March 2025 Page: 2193−2203
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i4.6510

[9] M. Alfaro-Contreras, A. Ríos-Vila, J. J. Valero-Mas, J. M. Iñesta, and J. Calvo-Zaragoza, "Decoupling music notation to improve end-to-end Optical Music Recognition," Pattern Recognit Lett, vol. 158, pp. 157–163, Jun. 2022, doi: 10.1016/j.patrec.2022.04.032.

[10] H. Dwiki Kahingide and A. Salam, "Deployment of Kidney Tumor Disease Object Detection Using CT-Scan with YOLOv5," Journal of Applied Informatics and Computing (JAIC), vol. 8, no. 1, pp. 98–105, Jul. 2024, [Online]. Available: http://jurnal.polibatam.ac.id/index.php/JAIC

[11] C. Garrido-Munoz, A. Rios-Vila, and J. Calvo-Zaragoza, "Proceedings of the 4th International Workshop on Reading Music Systems," in End-to-End Graph Prediction for Optical Music Recognition , Nov. 2022. doi: 10.48550/arXiv.2211.13285.

[12] P. Kania, D. Kania, and T. Łukaszewicz, "A hardware-oriented algorithm for real-time music key signature recognition," Applied Sciences (Switzerland), vol. 11, no. 18, Sep. 2021, doi: 10.3390/app11188753.

[13] H. Nakata and T. Nakanishi, "Music Impression Extraction Method by chord Impressions and Its Application to Music Media Retrieval," in Proceedings - 22nd IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing, SNPD 2021-Fall, Institute of Electrical and Electronics Engineers Inc., 2021, pp. 68–73. doi: 10.1109/SNPD51163.2021.9704990.

[14] A. Saxena et al., "Abnormal Health Monitoring and Assessment of a Three-Phase Induction Motor Using a Supervised CNN-RNN-Based Machine Learning Algorithm," Math Probl Eng, vol. 2023, 2023, doi: 10.1155/2023/1264345.

[15] Y. Liu, R. Wu, Y. Wu, L. Luo, and W. Xu, "A Stave-Aware Optical Music Recognition on Monophonic Scores for Camera-Based Scenarios," Applied Sciences (Switzerland), vol. 13, no. 16, Aug. 2023, doi: 10.3390/app13169360.

[16] A. Ríos-Vila, D. Rizo, J. M. Iñesta, and J. Calvo-Zaragoza, "End-to-end optical music recognition for pianoform sheet music," in International Journal on Document Analysis and Recognition, Springer Science and Business Media Deutschland GmbH, Sep. 2023, pp. 347–362. doi: 10.1007/s10032-023-00432-z.

[17] S. Edirisooriya, H.-W. Dong, J. McAuley, and T. Berg-Kirkpatrick, "An Empirical Evaluation of End-to-End Polyphonic Optical Music Recognition," Aug. 2021, [Online]. Available: http://arxiv.org/abs/2108.01769

[18] A. Ríos-Vila, M. Esplà-Gomis, D. Rizo, P. J. Ponce de León, and J. M. Iñesta, "Applying Automatic Translation for Optical Music Recognition's Encoding Step," Applied Sciences, vol. 11, no. 9, p. 3890, Apr. 2021, doi: 10.3390/app11093890.

[19] P. Torras, A. Barao, L. Kang, and A. Fornes, "Proceedings of the 4th International Workshop on Reading Music Systems," Improving Handwritten Music Recognition through Language Model Integration, Nov. 2022, doi: 10.48550/arXiv.2211.13285.

[20] Karsono, J. Daryanto, Rukayah, T. Budiharto, A. Yahya, and M. Anton Nugroho, "Musescore Software Training for the Development of TPACK-Based Music Learning in Elementary Schools," Dinamisia : Jurnal Pengabdian Kepada Masyarakat, vol. 7, no. 4, pp. 1128–1138, Aug. 2023, doi: 10.31849/dinamisia.v7i4.14807.

[21] Khawaja Tehseen Ahmed, N. Shahid, S. B. ud D. Tahir, A. Shabir, M. Y. Khan, and M. Hameed, "Signature Elevation Using Parametric Fusion for Large Convolutional Network for Image Extraction," VFAST Transactions on Software Engineering, vol. 12, no. 2, pp. 174–191, Jun. 2024, doi: 10.21015/vtse.v12i2.1810.

[22] H. Talebi and P. Milanfar, "Learning to Resize Images for Computer Vision Tasks," in 2021 IEEE/CVF International Conference on Computer Vision (ICCV), IEEE, Oct. 2021, pp. 487–496. doi: 10.1109/ICCV48922.2021.00055.

[23] R. Deléarde, C. Kurtz, P. Dejean, and L. Wendling, "Segment my object: A pipeline to extract segmented objects in images based on labels or bounding boxes," in VISIGRAPP 2021 - Proceedings of the 16th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications, SciTePress, 2021, pp. 618–625. doi: 10.5220/0010324006180625.

[24] T. Kumar, A. Mileo, R. Brennan, and M. Bendechache, "Image Data Augmentation Approaches: A Comprehensive Survey and Future directions," Jan. 2023, doi: 10.1109/ACCESS.2024.3470122.

[25] D. Say, S. Zidi, S. M. Qaisar, and M. Krichen, "Automated Categorization of Multiclass Welding Defects Using the X-ray Image Augmentation and Convolutional Neural Network," Sensors, vol. 23, no. 14, Jul. 2023, doi: 10.3390/s23146422.

[26] M. Alfaro-Contreras and J. J. Valero-Mas, "Exploiting the Two-Dimensional Nature of Agnostic Music Notation for Neural Optical Music Recognition," Applied Sciences, vol. 11, no. 8, p. 3621, Apr. 2021, doi: 10.3390/app11083621.

[27] B. Ait Skourt, A. El Hassani, and A. Majda, "Mixed-pooling-dropout for convolutional neural network regularization," Journal of King Saud University - Computer and Information Sciences, vol. 34, no. 8, pp. 4756–4762, Sep. 2022, doi: 10.1016/j.jksuci.2021.05.001.

[28] Z. Lyu et al., "Back-Propagation Neural Network Optimized by K-Fold Cross-Validation for Prediction of Torsional Strength of Reinforced Concrete Beam," Materials, vol. 15, no. 4, Feb. 2022, doi: 10.3390/ma15041477.

[29] M. Mamun, A. Farjana, M. Al Mamun, and M. S. Ahammed, "Lung cancer prediction model using ensemble learning techniques and a systematic review analysis," in 2022 IEEE World AI IoT Congress, AIIoT 2022, Institute of Electrical and Electronics Engineers Inc., 2022, pp. 187–193. doi: 10.1109/AIIoT54504.2022.9817326.

[30] D. S. Soper, "Greed is good: Rapid hyperparameter optimization and model selection using greedy k-fold cross validation," Electronics (Switzerland), vol. 10, no. 16, Aug. 2021, doi: 10.3390/electronics10161973.

[31] R. Artanto, W. Sujana, I. Made, and A. Agastya, "Application of Machine Learning Algorithm for Osteoporosis Disease Prediction System," Journal of Applied Informatics and Computing (JAIC), vol. 8, no. 2, p. 304, Dec. 2024, [Online]. Available: http://jurnal.polibatam.ac.id/index.php/JAIC

[32] T. Safa Nabila and A. Salam, "Classification of Brain Tumors by Using a Hybrid CNN-SVM Model," Journal of Applied Informatics and Computing (JAIC), vol. 8, no. 2, p. 241, Dec. 2024, [Online]. Available: http://jurnal.polibatam.ac.id/index.php/JAIC

[33] M. Heydarian, T. E. Doyle, and R. Samavi, "MLCM: Multi-Label Confusion Matrix," IEEE Access, vol. 10, pp. 19083–19095, 2022, doi: 10.1109/ACCESS.2022.3151048.

[34] J. Qu, C. Song, J. Bai, G. Feng, X. Shi, and J. Ma, "A Machine-Learning-Based Method for Identifying the Failure Risk State of Fissured Sandstone under Water–Rock Interaction," Sensors, vol. 24, no. 17, Sep. 2024, doi: 10.3390/s24175752.

[35] R. R. Adhitya, Wina Witanti, and Rezki Yuniarti, "Perbandingan Metode Cart Dan Naïve Bayes Untuk Klasifikasi Customer Churn," INFOTECH journal, vol. 9, no. 2, pp. 307–318, Jul. 2023, doi: 10.31949/infotech.v9i2.5641.