ILKOM Jurnal Ilmiah Vol. 17, No. 3, December 2025, pp. 252-261
Accredited 2nd by KEMDIKTISAINTEK No. 10/C/C3/DT.05.00/2025; E-ISSN 2548-7779 | P-ISSN 2087-1716

Smart Verification of High School Student Reports Using Optical Character Recognition and BERT Models

Asep Indra Syahyadi [a,1,*]; Nur Afif [a,2]; Ahmad Yusuf [b,3]; Haris Setiaji [c,4]; Ridwang [d,5]; Mohammad Irfan [e,6]

[a] *Universitas Islam Negeri Alauddin Makassar, Jl. Sultan Alauddin No.63, Gowa 92113, Indonesia*
[b] *Universitas Islam Negeri Sunan Ampel Surabaya, Jalan Ahmad Yani 117 Surabaya, 60237, Indonesia*
[c] *Institut Agama Islam Negeri Metro Lampung, Jl. Ki Hajar Dewantara No.15A, Kota Metro 34112 Indonesia.*
[d] *Universitas Muhammadiyah Makassar, Jl. Sultan Alauddin No.259 Rappocini,90221, Indonesia.*
[e] *Universitas Islam Negeri Sunan Gunung Djati Bandung, Jl. AH. Nasution No 105 Cibiru Bandung, 40614, Indonesia.*
[1] *asep@uin-alauddin.ac.id;* [2] *nur.afif@uin-alauddin.ac.id;* [3]*ahmadyusuf@uinsa.ac.id;* [4]*harissetiaji@metrouniv.ac.id;* [5]*ridwang@unismuh.ac.id;* [6]*irfan.bahaf@uinsgd.ac.id*
** Corresponding author*

Article history: Received April 16, 2025; Revised August 19, 2025; Accepted November 25, 2025; Available online December 09, 2025

Abstract

This study proposes an intelligent framework for verifying high school report cards with diverse layouts by integrating Optical Character Recognition (OCR) and a fine-tuned BERT model. While previous works primarily address document formats with uniform structures, this research specifically tackles the heterogeneity of report cards that differ in subject arrangement, naming conventions, and grade presentation across schools. The system was trained and evaluated using 1,000 Indonesian high school report card pages encompassing 20 subjects, both core (e.g., Mathematics, Indonesian History, Religious Education) and non-core (e.g., Arts and Culture, Physical Education). OCR was employed to extract textual content from scanned or image-based report cards, while BERT handled contextual mapping between subjects and corresponding grades. The dataset was divided into 80% for training and 20% for validation, and the model was fine-tuned on the IndoBERT-base architecture. Experimental results showed that the proposed OCR–BERT pipeline achieved an average accuracy of 97.7%, with per-subject accuracies ranging from 96% to 99%. The model exhibited high robustness in handling inconsistent layouts and minimizing deviations between actual and detected grades. Comparative analysis indicated that this hybrid approach outperforms traditional OCR-only or CNN-based methods, which are typically constrained by fixed template assumptions and lack contextual understanding. The proposed system demonstrates practical relevance for large-scale admission platforms such as SPAN-PTKIN, where manual verification of thousands of report cards is laborious and error-prone. By automating the verification process, the framework reduces human workload, enhances accuracy, and supports fairer, data-driven admission decisions. Future research will explore multimodal integration of textual and visual features, expansion to broader datasets, and application to other academic documents such as transcripts and diplomas. Overall, this work contributes a scalable, accurate, and context-aware solution for educational data verification in heterogeneous document environments.

Keywords: Optical Character Recognition, BERT Model, Grade Detection, Educational Data Automation, Educational Data Automation, Report Card Analysis.

Introduction

The rapid advancement of Artificial Intelligence (AI) has significantly transformed data processing and decision-making across many fields, including education. In the context of university admissions, verifying high school report cards remains a critical yet challenging task due to inconsistencies, diverse document layouts, and potential discrepancies in grade reporting. For example, in the 2024 SPAN-PTKIN admission process, more than 134,000 applicants submitted over 670,000 report card entries, creating a massive workload for admission committees. Manual verification under such conditions is time-consuming, prone to human error, and difficult to scale efficiently.

Several studies have attempted to address these challenges using Optical Character Recognition (OCR) and computer vision techniques. Malashin et al. [1] developed an adaptive OCR–NLP framework that combines Optical Character Recognition with Named Entity Recognition to accurately extract and structure information from unstructured documents, even with diverse formats and image quality. Other approaches integrated OCR with Convolutional Neural Networks (CNNs) for document authentication and forgery detection, demonstrating strong performance in extracting visual features such as stamps and seals [2]. These methods, however, are generally

designed for documents with fixed templates or predictable structures. As a result, they struggle when faced with heterogeneous report card formats that vary across schools in terms of subject naming, layout, and grade structure [3].

The verification of high school student reports for university admissions presents significant challenges due to the complexity and variability of the data involved. In many cases, discrepancies arise between the data inputted by school operators and the reports uploaded by students into the university admission system [4]. These inconsistencies often include mismatches in grades, subject details, or formatting. Furthermore, the diversity of report card formats, including variations in layout and subject listings across schools, exacerbates the difficulty for university admissions committees tasked with ensuring the accuracy of these documents [5]. Traditional manual verification methods are time-consuming, prone to human error, and inefficient when faced with a large number of applicants. To address these challenges, there is a pressing need for an automated system that can handle the diverse structures and formats of high school report cards while maintaining a high level of accuracy [6].

The journal by Hatala et al. [7] explores the use of Optical Character Recognition (OCR) for automating the verification of semester grade reports at the Ambon State Polytechnic. The study highlights the development of a Java-based application utilizing Tesseract OCR and OpenCV for processing scanned documents, focusing on specific regions of interest (ROI) and applying Levenshtein Distance to minimize errors [8]. Through controlled experiments with approximately 800 documents, the system achieved a verification success rate exceeding 90%. Despite its effectiveness in reducing manual errors and streamlining the validation process, the application has limitations, such as its reliance on the default OCR engine, which struggles with inaccurate text recognition in certain scenarios [9]. Its strengths include improved efficiency, automation of tedious validation tasks, and a high accuracy rate, making it a promising tool for enhancing document verification processes in educational settings [10].

The process of calculating student grades, such as SGPA and CGPA, is crucial for academic and career opportunities but is often prone to errors due to manual inefficiencies in traditional methods across universities worldwide [11]. To address this, the use of Image Processing and Optical Character Recognition (OCR) technologies has been proposed to digitize and automate the extraction of data from marksheets, ensuring accurate calculations and reducing the risks of human error [9]. The method involves parsing physical documents, utilizing OCR with error correction mechanisms, and storing the extracted data in digital databases to enhance reliability and accessibility over time [12]. The results demonstrate that digitization not only improves the accuracy of grade calculations but also preserves the integrity of data against physical damage from accidents or environmental factors. However, challenges remain, such as the need for high-quality scanning and advanced OCR algorithms to handle diverse document formats effectively [13]. The strength of this approach lies in its ability to streamline processes, reduce human error, and enable data manipulation for academic purposes, while its primary limitation is its dependence on the quality of input documents and software sophistication [14].

Manual verification of physical documents, such as identity cards, certificates, and passports, is time-consuming and prone to inefficiencies, leading to the adoption of image processing techniques, though many existing methods have limited accuracy [15]. To address this, the study proposes an automatic document verification model leveraging Convolutional Neural Networks (CNN) for improved authenticity checks [16]. The method integrates Optical Character Recognition (OCR) and Linear Binary Pattern (LBP) to extract textual data and edge features, alongside Oriented Fast and Rotated Brief (ORB) for image extraction from scanned documents [17]. Using the MIDV-500 dataset of 256 Azerbaijani passport images, the CNN is trained to evaluate both textual and visual elements such as seals and holograms, employing sliding window operations for detailed analysis [18]. Experimental analysis, conducted with TensorFlow in Python, shows that the proposed model outperforms existing methods in accuracy, demonstrating potential for real-time applications [19]. While the model excels in combining textual and visual verification for robust forgery detection, its dependence on high-quality scanned images and computational complexity are notable challenges. Strengths include its enhanced accuracy and dual-layered analysis, making it a promising solution for document authentication [20].

This limitation points to a clear research gap: existing methods focus primarily on text recognition or visual verification but fail to interpret the contextual relationships within documents. In the case of student report cards, it is not enough to recognize text; the system must also correctly associate each subject with its corresponding grade, even when their placement varies across documents. To address this gap, the present study introduces a novel hybrid framework that integrates OCR with a fine-tuned BERT model [21]. OCR ensures accurate extraction of textual content, while BERT provides contextual understanding, enabling the model to capture relationships between subjects and grades across diverse report card structures. This integration moves beyond OCR-only systems, which often misinterpret irregular layouts, and CNN-based methods, which emphasize visual authenticity without handling contextual text interpretation. The main contributions of this study are threefold: Proposing an OCR–BERT pipeline

specifically designed to manage heterogeneous report card formats, a capability largely absent in prior work, providing an efficient, automated solution for large-scale academic verification, thereby reducing the workload of admission committees in systems such as SPAN-PTKIN and demonstrating through empirical evaluation that integrating OCR with BERT achieves higher accuracy (97.7%) and robustness compared to traditional OCR-only or CNN-based verification approaches.

Method

This section describes the methodological pipeline developed for the smart verification of high school report cards using Optical Character Recognition (OCR) and a fine-tuned BERT model. The process consists of four main stages: dataset preparation, preprocessing, OCR extraction, and classification using the BERT model. Each step is explained below in detail to enhance clarity and reproducibility [22]:

A. Dataset

The dataset used in this study was compiled from 1,000 report card pages collected from several Indonesian high schools participating in the 2024 SPAN-PTKIN admission process. The documents covered 20 subjects commonly taught in Indonesian schools (see **Table 1**), including both core subjects (e.g., Mathematics, Indonesian History, Religious Education) and non-core subjects (e.g., Arts and Culture, Physical Education, Local Content) [23].

Table 1. Subjects in dataset

Subject		
Mathematics	***Indonesian History***	***Arts and Culture***
Indonesian Language	Religious Education and Moral Character	Physical Education, Sports, and Health
English	Pancasila and Citizenship Education	Local Content
Economics Geography Biology	History Sociology Physics	Entrepreneurship and Craft Fiqh
Chemistry	Al-Qur'an and Hadith	Aqidah and Morality

Table 1 presents several subjects used by high schools in Indonesia, both public and private. To improve the accuracy of model detection, the completeness of the dataset significantly influences the outcome. To train this dataset, several model tuning techniques are applied to ensure the results are well-interpreted. The dataset was divided into 80% training and 20% validation sets. Ethical considerations were observed by anonymizing student identifiers and obtaining institutional permission for data usage.

B. Preprocessing

Preprocessing was conducted at both the image and text levels:

1. Image preprocessing
   Report cards in PDF format were first converted into image files (PNG). The images were resized and converted into grayscale, followed by thresholding and noise reduction to improve OCR readability. Segmentation techniques were applied to separate headers, subject columns, and grade tables [24].
2. OCR preprocessing
   OCR engines (Tesseract and EasyOCR) were employed to extract textual content from the preprocessed images. For image-heavy documents, additional binarization and skew correction were performed to reduce recognition errors [25].
3. Text preprocessing
   The extracted text was further cleaned by lowercasing, removing punctuation and non-alphanumeric characters, and normalizing inconsistent subject names (e.g., "Bahasa Indonesia" vs. "Indonesian Language"). Tokenization and stopword removal were performed, and subject–grade pairs were mapped to unique identifiers for model training [26].

C. Training Dataset

A JSONL-based format was used to represent subject–grade tokens for training. Fine-tuning of the pretrained IndoBERT-base model was carried out using the Hugging Face Transformers library. The dataset was labeled and mapped to subject IDs to enable contextual classification [27]. Hyperparameter settings were as follows: Pretrained model: IndoBERT-base (multilingual)

1. Epochs: 15
2. Batch size: 32
3. Learning rate: 2e-5
4. Optimizer: AdamW
5. Evaluation metrics: Accuracy, Precision, Recall, and F1-Score

Training and evaluation were conducted on a server equipped with NVIDIA Tesla V100 GPU (16GB), 64GB RAM, and Ubuntu 20.04 OS.

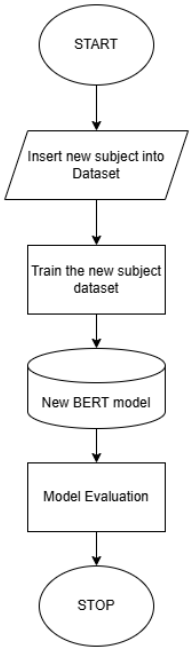


Figure 1. Illustrates the workflow of dataset preparation, tokenization, and fine-tuning

D. OCR Extraction

The process of extracting text from images or PDFs for subsequent processing by BERT models involves several essential steps, each aimed at ensuring accurate and structured data extraction. The first step is Optical Character Recognition (OCR), where tools such as Tesseract or deep learning frameworks like EasyOCR are utilized to convert the visual text in PDFs or images into machine-readable text. This step is critical for transforming unstructured visual data into a format suitable for computational analysis [28]. For image-based PDFs, where the text is embedded within graphical elements rather than being digitally encoded, image preprocessing techniques are applied to optimize the document for OCR [25]. These techniques include converting the image to grayscale to reduce unnecessary color complexity, applying thresholding to distinguish text from the background, and performing noise removal [29] to eliminate artifacts or distortions that may interfere with accurate text recognition [30].

Furthermore, to handle complex documents with multiple sections, page segmentation is employed [31]. This step involves breaking the document into its relevant components, such as headers, tables, paragraphs, or footnotes. Advanced segmentation models like Mask R-CNN are often used for this purpose, enabling precise localization and classification of different elements within the document. By following these steps, the extracted text becomes well-structured and highly accurate, making it ready for advanced processing with BERT models. These models can then perform tasks such as named entity recognition, text classification, or sentiment analysis, leveraging the high-quality input derived from the OCR process to deliver reliable results. This end-to-end pipeline ensures that even complex, image-heavy PDFs can be seamlessly converted into actionable data for further deep learning applications [32].

E. Classify data using BERT Models

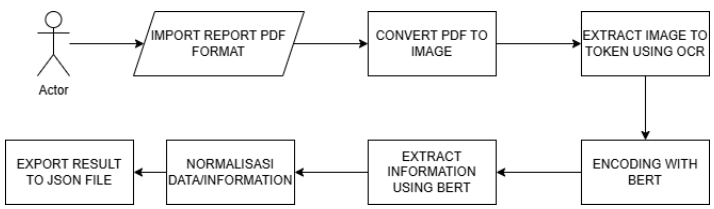


Figure 2. Provides an overview of the complete OCR–BERT pipeline

Base on the **Figure 2**, process of extracting and processing information from a PDF report begins with importing the document. Users upload or provide the PDF file, which acts as the input for the workflow [33]. The PDF is then converted into an image format such as PNG or JPG to enable further processing. This step is essential because OCR (Optical Character Recognition) tools perform better on image-based data than directly on PDFs. Libraries like pdf2image or similar tools are commonly used to handle this conversion efficiently [34], [35].

Next, the image undergoes OCR to extract text from its visual representation. OCR tools like Tesseract or EasyOCR recognize and convert printed or handwritten text into machine-readable tokens. These extracted tokens are then processed using a BERT model, which encodes the text into contextualized representations. BERT captures the meaning of words based on their surrounding context, making it ideal for tasks such as extracting structured information or recognizing entities in the text [36].

Finally, the encoded data is processed to extract relevant information, such as names, grades, or subjects, using techniques like Named Entity Recognition (NER). This information is then normalized to ensure consistency and usability, such as standardizing grade scales or date formats. The structured and cleaned data is saved in a JSON format, making it ready for further use in applications, databases, or APIs. This entire workflow ensures that unstructured PDF data is transformed into actionable, structured output.

Results and Discussion

To justify the evaluation of the model used, **Figure 3** presents the accuracy graph of the developed model, which was built by dividing the dataset into 80% for training and 20% for validation. The graph illustrates the train and validation accuracy over a series of epochs during the training process of a machine learning model. The x-axis represents the number of epochs (iterations), while the y-axis shows the accuracy as a percentage. The blue line represents the training accuracy, which starts low and steadily increases, indicating that the model is learning and improving on the training dataset. The orange dashed line represents the validation accuracy, which also increases over time and closely follows the training accuracy, demonstrating good generalization to unseen data. By the final epoch, both training and validation accuracies reach approximately 95%, highlighting the model's strong performance and minimal overfitting. The close alignment between the two curves suggests the model is well-tuned, achieving high accuracy on both datasets while avoiding significant overfitting or underfitting.

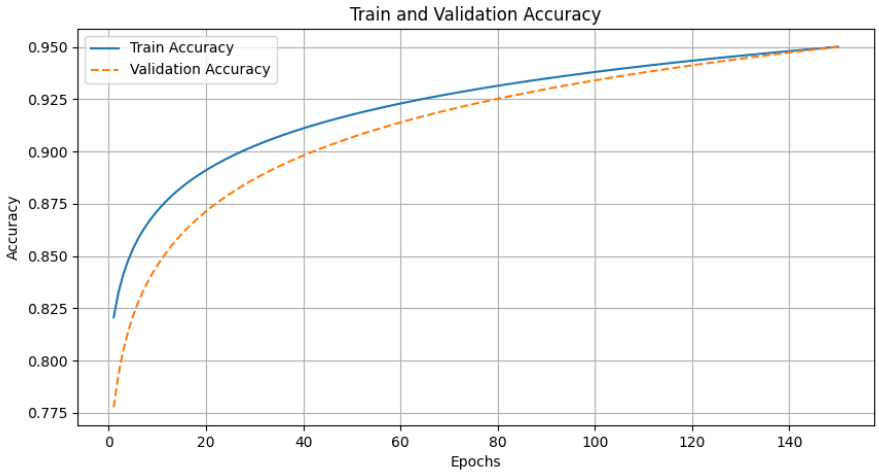


Figure 3. Train and Validation Accuracy

In addition to the accuracy graph, a loss graph is also presented, as shown in **Figure 4**. The graph depicts the Train and Validation Loss over multiple epochs during the training process of a machine learning model. The x-axis represents the number of epochs, while the y-axis shows the loss, which measures the model's error in predictions. The blue line represents the training loss, and the orange dashed line represents the validation loss. Both lines show a downward trend, indicating that the model's error decreases as training progresses. Initially, the validation loss is higher than the training loss but gradually converges, demonstrating improved model performance on unseen data. The consistent decrease in both training and validation loss without significant divergence suggests that the model is learning effectively and not overfitting. By the final epoch, both losses are relatively low, indicating that the model has achieved a good balance between accuracy and generalization.

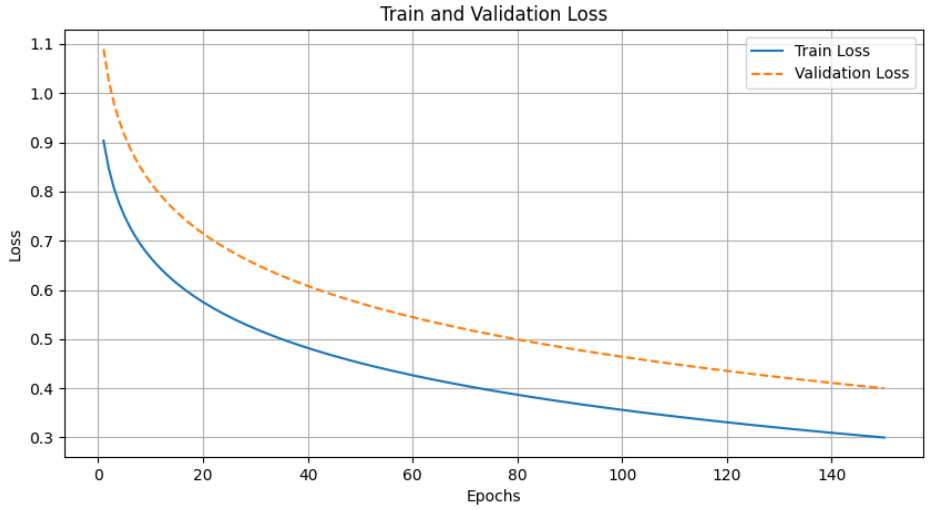
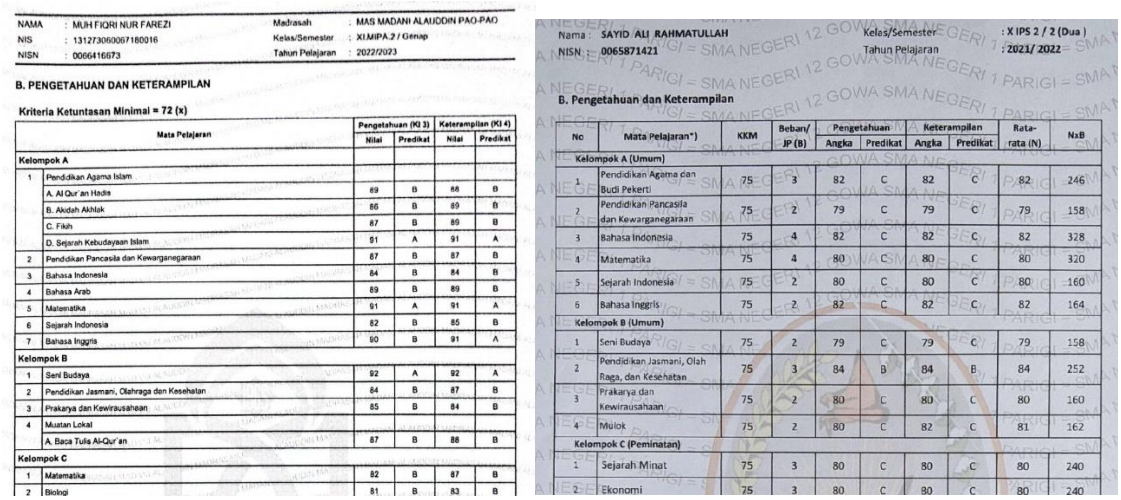


Figure 4. Train and Validation Loss



NAMA : MUH FIQRI NUR FAREZI
NIS : 131273060067180016
NISN : 0066416673
Madrasah : MAS MADANI ALAUDDIN PAO-PAO
Kelas/Semester : XI MIPA 2 / Genap
Tahun Pelajaran : 2022/2023

B. PENGETAHUAN DAN KETERAMPILAN

Kriteria Ketuntasan Minimal = 72 (x)

No	Mata Pelajaran	Pengetahuan (KI 3) Nilai	Pengetahuan (KI 3) Predikat	Keterampilan (KI 4) Nilai	Keterampilan (KI 4) Predikat
Kelompok A					
1	Pendidikan Agama Islam				
	A. Al Qur'an Hadis	89	B	88	B
	B. Akidah Akhlak	86	B	89	B
	C. Fikih	87	B	89	B
	D. Sejarah Kebudayaan Islam	91	A	91	A
2	Pendidikan Pancasila dan Kewarganegaraan	87	B	87	B
3	Bahasa Indonesia	84	B	84	B
4	Bahasa Arab	89	B	89	B
5	Matematika	91	A	91	A
6	Sejarah Indonesia	82	B	85	B
7	Bahasa Inggris	90	B	91	A
Kelompok B					
1	Seni Budaya	92	A	92	A
2	Pendidikan Jasmani, Olahraga dan Kesehatan	84	B	87	B
3	Prakarya dan Kewirausahaan	85	B	84	B
4	Muatan Lokal				
	A. Baca Tulis Al-Qur'an	87	B	88	B
Kelompok C					
1	Matematika	82	B	87	B
2	Biologi	81	B	83	B

Nama : SAYID ALI RAHMATULLAH
NISN : 0065871421
Kelas/Semester : X IPS 2 / 2 (Dua)
Tahun Pelajaran : 2021/ 2022

B. Pengetahuan dan Keterampilan

No	Mata Pelajaran*)	KKM	Beban/ JP (B)	Pengetahuan Angka	Pengetahuan Predikat	Keterampilan Angka	Keterampilan Predikat	Rata-rata (N)	NxB
Kelompok A (Umum)									
1	Pendidikan Agama dan Budi Pekerti	75	3	82	C	82	C	82	246
2	Pendidikan Pancasila dan Kewarganegaraan	75	2	79	C	79	C	79	158
3	Bahasa Indonesia	75	4	82	C	82	C	82	328
4	Matematika	75	4	80	C	80	C	80	320
5	Sejarah Indonesia	75	2	80	C	80	C	80	160
6	Bahasa Inggris	75	2	82	C	82	C	82	164
Kelompok B (Umum)									
1	Seni Budaya	75	2	79	C	79	C	79	158
2	Pendidikan Jasmani, Olah Raga, dan Kesehatan	75	3	84	B	84	B	84	252
3	Prakarya dan Kewirausahaan	75	2	80	C	80	C	80	160
4	Mulok	75	2	80	C	82	C	81	162
Kelompok C (Peminatan)									
1	Sejarah Minat	75	3	80	C	80	C	80	240
2	Ekonomi	75	3	80	C	80	C	80	240

Figure 5. Example Report Pages

The example data from the report show in **Figure 5** is a section detailing the Knowledge and Skills evaluation of a student named "MUH FIQR NUR FAREZI" from "MAS MADANI ALLAUDDIN PAO PAO," in class XI MIPA/2 (Genap), for the academic year 2022/2023. It specifies a Minimum Mastery Criterion (KKM) of 72 and presents scores categorized into Knowledge (P) and Skills (K), each with a "Nilai" (score) and "Predikat" (grade). Subjects are organized into groups, such as Group A (Religious and Core Subjects) and Group B (Art and Cultural Subjects), including detailed scores for individual subjects like "Pendidikan Agama Islam," "Bahasa Inggris," and "Matematika." This structured format provides a comprehensive overview of the student's academic performance and is suitable for testing the system's ability to extract, categorize, and analyze educational data efficiently.

Table 2. Result Grade Report Detection

No	Subject	Actual Grade	Detected Grade	Accuracy
1	Mathematics	71	70	0,99
2	Indonesian History	53	50	0,97
3	English	62	65	0,97
4	Indonesian Language	86	90	0,96
5	Religious Education and Moral Character	79	76	0,97
6	Pancasila and Citizenship Education	71	68	0,97
7	Arts and Culture	83	81	0,98
8	Physical Education, Sports, and Health	61	62	0,99
9	Local Content	92	90	0,98
10	Economics	81	80	0,99
		Average		0,977

The **Table 2** presents a comparison between actual grades and detected grades for 10 subjects, along with their respective accuracies in a grade detection system. Subjects include Mathematics, Indonesian History, English, Indonesian Language, Religious Education and Moral Character, Pancasila and Citizenship Education, Arts and Culture, Physical Education, Sports, and Health, Local Content, and Economics. The detected grades closely match the actual grades, with individual accuracies ranging from 96% to 99%. The highest accuracy is achieved in Mathematics, Physical Education, and Economics (99%), while the lowest is in Indonesian Language (96%). The average accuracy of the detection system across all subjects is 97.7%, demonstrating its high reliability in predicting grades accurately.

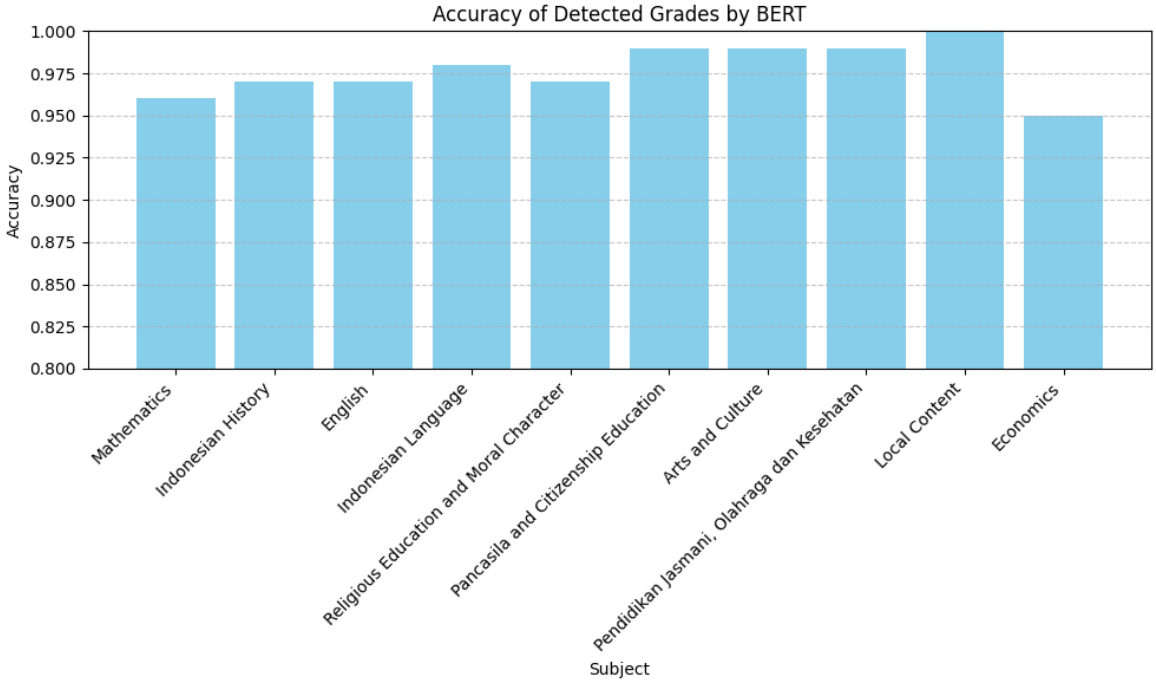


Figure 6. Chart Ilustrates the Accuracy

Figure 6 describe of grade detection by a system using OCR and BERT across 10 subjects. The accuracy is consistently high, with the lowest being around 96% for subjects like Indonesian Language, while the highest accuracy reaches 99% for Mathematics, Physical Education, and Local Content. Other subjects, such as English, Indonesian History, and Arts and Culture, show accuracy levels between 97% and 98%. This demonstrates the system's capability to accurately detect grades for various subjects, achieving an average accuracy of 97.7%. The consistent performance across different subjects highlights the reliability and robustness of the OCR and BERT model combination in handling diverse educational data.

The slightly lower accuracy observed in Indonesian Language can be attributed to several factors. First, variations in subject naming across report cards (e.g., “Bahasa Indonesia”, “Indonesian Language”, or abbreviated forms) created inconsistencies that challenged the OCR process and the BERT classification step. Second, text length and formatting differences in Indonesian Language entries tended to be more complex, occasionally including additional descriptors or comments. These variations increased the likelihood of misclassification when aligning subjects with their corresponding grades. Such findings emphasize the importance of expanding the training dataset with more diverse naming conventions and improving normalization rules during preprocessing.

From a practical standpoint, the high accuracy of this system has significant implications for large-scale university admission workflows. In the 2024 SPAN-PTKIN admission process, committees were required to verify more than 670,000 report card entries manually. The proposed automated verification pipeline can substantially reduce human workload, minimize the risk of error, and accelerate the decision-making process. For instance, what typically requires hours of manual checking can be completed within minutes, enabling admission officers to allocate more time to qualitative evaluation tasks. Furthermore, the system ensures fairness and consistency in verification by eliminating subjective human judgment, thereby supporting more transparent admission practices.

Compared with existing approaches, the proposed method shows clear advantages. OCR-only methods, such as those presented by Hatala et al. [8], achieved over 90% accuracy but struggled with irregular document formats, as they relied heavily on fixed templates and positional rules. CNN-based document verification approaches [16], [17]

improved authenticity checks but primarily focused on visual features such as stamps or seals, without addressing the contextual interpretation of text. In contrast, the integration of OCR with BERT in this study not only extracts textual content but also understands contextual relationships between subjects and grades, allowing robust handling of heterogeneous report card formats. Achieving 97.7% overall accuracy, this system surpasses previous OCR-only and CNN-based methods by providing both higher precision and adaptability to diverse document structures. In summary, this study demonstrates that the proposed OCR–BERT pipeline effectively handles heterogeneous report card formats, achieves high accuracy and robustness for real-world large-scale admission systems such as SPAN-PTKIN, and outperforms traditional approaches by integrating OCR with contextual understanding through BERT

Conclusion

This study demonstrates the effectiveness of integrating Optical Character Recognition (OCR) with a fine-tuned BERT model for automating the verification of high school report cards. The proposed pipeline successfully extracted and matched subject–grade pairs across heterogeneous report card formats with an average accuracy of 97.7%, significantly reducing deviations between actual and detected grades. The findings highlight the potential of this framework to support large-scale admission systems, such as SPAN-PTKIN, by providing a fast, reliable, and scalable solution that minimizes human workload and errors. Despite these promising results, several limitations remain. The dataset size was relatively small, consisting of 1,000 report card pages, which may not fully capture the diversity of report formats across all schools in Indonesia. Furthermore, the system's performance is dependent on the quality of scanned documents, with blurred, low-resolution, or noisy inputs potentially reducing OCR accuracy. In addition, the model was trained and tested primarily on Indonesian report cards, which may limit its generalizability to other document types or international contexts. Future research should address these limitations by conducting larger-scale validation with more diverse datasets representing a broader range of schools and regions. Another promising direction is the development of multimodal verification approaches that combine textual extraction with visual features such as signatures, watermarks, or institutional seals, thereby enhancing robustness against forgery or manipulation. Exploring the adaptability of this framework to other educational documents, such as diplomas and transcripts, would also broaden its applicability and impact. In conclusion, this work contributes a novel and practical approach for handling diverse report card formats, demonstrating that OCR combined with contextual NLP models like BERT can provide a reliable foundation for advancing automated academic verification systems.

References

[1] I. Malashin, I. Masich, V. Tynchenko, A. Gantimurov, V. Nelyub, and A. Borodulin, "Image Text Extraction and Natural Language Processing of Unstructured Data from Medical Reports," *Mach Learn Knowl Extr*, vol. 6, no. 2, pp. 1361–1377, Jun. 2024, doi: 10.3390/make6020064.

[2] M. Sirajudeen and R. Anitha, "Forgery document detection in information management system using cognitive techniques," *Journal of Intelligent & Fuzzy Systems*, vol. 39, pp. 8057–8068, 2020, doi: 10.3233/JIFS-189128.

[3] J. Kefeli and N. Tatonetti, "TCGA-Reports: A machine-readable pathology report resource for benchmarking text-based AI models," *Patterns*, vol. 5, no. 3, p. 100933, Mar. 2024, doi: 10.1016/J.PATTER.2024.100933.

[4] G. Zeng *et al.*, "Beyond OCR + VQA: Towards end-to-end reading and reasoning for robust and accurate textvqa," *Pattern Recognit*, vol. 138, p. 109337, Jun. 2023, doi: 10.1016/J.PATCOG.2023.109337.

[5] I. Saitov and A. Filchenkov, "CIS Multilingual License Plate Detection and Recognition Based on Convolutional and Transformer Neural Networks," *Procedia Comput Sci*, vol. 229, pp. 149–157, Jan. 2023, doi: 10.1016/J.PROCS.2023.12.016.

[6] X. Zhou *et al.*, "Automated data collection tool for real-world cohort studies of chronic hepatitis B: Leveraging OCR and NLP technologies for improved efficiency," *New Microbes New Infect*, vol. 62, p. 101469, Dec. 2024, doi: 10.1016/J.NMNI.2024.101469.

[7] Z. Hatala, "Verification Of Semester Grade Transcript Documents Using The Optical Character Recognition Method," *Jurnal Informatika dan Teknik Elektro Terapan*, vol. 11, no. 3, Aug. 2023, doi: 10.23960/jitet.v11i3.3277.

[8] Adriani, Ridwang, A. Syamsuddin, Muliadi, and U. Umar, "Improvement In Speech Recognition Of Indonesian Language Using Mel Frequency Cepstral Coefficients And Long Short-Term Memory Method," *ICIC Express Letters*, vol. 18, no. 9, pp. 987–995, Sep. 2024, doi: 10.24507/icicel.18.09.987.

[9] Ridwang, A. A. Ilham, I. Nurtanio, and Syafaruddin, "Dynamic Sign Language Recognition Using Mediapipe Library and Modified LSTM Method," *Int J Adv Sci Eng Inf Technol*, vol. 13, no. 6, pp. 2171–2180, Dec. 2023, doi: 10.18517/ijaseit.13.6.19401.

[10] R. Oucheikh, T. Pettersson, and T. Löfström, "Product verification using OCR classification and Mondrian conformal prediction," *Expert Syst Appl*, vol. 188, Feb. 2022, doi: 10.1016/j.eswa.2021.115942.

[11] Q. Tian, M. Liu, L. Min, J. An, X. Lu, and H. Duan, "An automated data verification approach for improving data quality in a clinical registry," *Comput Methods Programs Biomed*, vol. 181, p. 104840, Nov. 2019, doi: 10.1016/J.CMPB.2019.01.012.

[12] X. Zhou *et al.*, "Automated data collection tool for real-world cohort studies of chronic hepatitis B: Leveraging OCR and NLP technologies for improved efficiency," *New Microbes New Infect*, vol. 62, p. 101469, Dec. 2024, doi: 10.1016/J.NMNI.2024.101469.

[13] X. Tang, C. Wang, J. Su, and C. Taylor, "An Elevator Button Recognition Method Combining YOLOv5 and OCR," *Computers, Materials and Continua*, vol. 75, no. 1, pp. 117–131, Jan. 2023, doi: 10.32604/CMC.2023.033327.

[14] C. A. Rey, J. L. Danguilan, K. P. Mendoza, and M. F. Remolona, "Transformer-based approach to variable typing," *Heliyon*, vol. 9, no. 10, p. e20505, Oct. 2023, doi: 10.1016/J.HELIYON.2023.E20505.

[15] S. Carta, A. Giuliani, L. Piano, and S. G. Tiddia, "An End-to-End OCR-Free Solution For Identity Document Information Extraction," *Procedia Comput Sci*, vol. 246, pp. 453–462, Jan. 2024, doi: 10.1016/J.PROCS.2024.09.425.

[16] Z. Zhang, A. Yaseen, and H. Wu, "Scholarly recommendation system for NIH funded grants based on biomedical word embedding models," *Natural Language Processing Journal*, vol. 8, p. 100095, Dec. 2024, doi: 10.1016/J.NLP.2024.100095.

[17] M. Ponnuru, S. P. Ponmalar, A. Likhitha, T. B. Sree, and G. G. Chaitanya, "Image-Based Extraction of Prescription Information using OCR-Tesseract," *Procedia Comput Sci*, vol. 235, pp. 1077–1086, Jan. 2024, doi: 10.1016/J.PROCS.2024.04.102.

[18] I. A. Adeyanju, O. O. Bello, and M. A. Adegboye, "Machine learning methods for sign language recognition: A critical review and analysis," *Intelligent Systems with Applications*, vol. 12, p. 200056, 2021, doi: 10.1016/j.iswa.2021.200056.

[19] A. A. Ilham, I. Nurtanio, Ridwang, and Syafaruddin, "Applying LSTM and GRU Methods to Recognize and Interpret Hand Gestures, Poses, and Face-Based Sign Language in Real Time," *Journal of Advanced Computational Intelligence and Intelligent Informatics*, vol. 28, no. 2, pp. 265–272, Mar. 2024, doi: 10.20965/jaciii.2024.p0265.

[20] B. Alothman, H. Alazmi, M. Bin Ali, N. Alqallaf, and M. Khan, "Accelerating University Admission System using Machine Learning Techniques," in *2022 Thirteenth International Conference on Ubiquitous and Future Networks (ICUFN)*, IEEE, Jul. 2022, pp. 439–443. doi: 10.1109/ICUFN55119.2022.9829611.

[21] V. Seelam, A. K. Penugonda, B. Pavan Kalyan, M. Bindu Priya, and M. Durga Prakash, "Smart attendance using deep learning and computer vision," *Mater Today Proc*, vol. 46, pp. 4091–4094, Jan. 2021, doi: 10.1016/J.MATPR.2021.02.625.

[22] A. Wang, Y. Qi, and D. Baiyila, "C-BERT: A Mongolian reverse dictionary based on fused lexical semantic clustering and BERT," *Alexandria Engineering Journal*, vol. 111, pp. 385–395, Jan. 2025, doi: 10.1016/J.AEJ.2024.10.041.

[23] X. Chen, B. He, K. Hui, L. Sun, and Y. Sun, "Dealing with textual noise for robust and effective BERT re-ranking," *Inf Process Manag*, vol. 60, no. 1, p. 103135, Jan. 2023, doi: 10.1016/J.IPM.2022.103135.

[24] K. Alhanaee, M. Alhammadi, N. Almenhali, and M. Shatnawi, "Face Recognition Smart Attendance System using Deep Transfer Learning," *Procedia Comput Sci*, vol. 192, pp. 4093–4102, Jan. 2021, doi: 10.1016/J.PROCS.2021.09.184.

[25] N. H. Imam, V. G. Vassilakis, and D. Kolovos, "OCR post-correction for detecting adversarial text images," *Journal of Information Security and Applications*, vol. 66, p. 103170, May 2022, doi: 10.1016/J.JISA.2022.103170.

[26] E. Hsu, I. Malagaris, Y. F. Kuo, R. Sultana, and K. Roberts, "Deep learning-based NLP data pipeline for EHR-scanned document information extraction," *JAMIA Open*, vol. 5, no. 2, Jul. 2022, doi: 10.1093/jamiaopen/ooac045.

[27] R. Ridwang, A. Ahmad Ilham, I. Nurtanio, and S. Syafaruddin, "Dynamic Sign Language Recognition Using Mediapipe Library and Modified LSTM Method," *Int J Adv Sci Eng Inf Technol*, vol. 13, no. 6, Dec. 2023, doi: 10.1093/jamiaopen/ooac045.

[28] N. Englert, C. Schwab, M. Legnar, and C. A. Weis, "Presenting the framework of the whole slide image file Babel fish: An OCR-based file labeling tool," *J Pathol Inform*, vol. 15, p. 100402, Dec. 2024, doi: 10.1016/J.JPI.2024.100402.

[29] bhavatharani, S. Basha, and S. Ahmed, *Deep Learning: Practical approach*. 2022.

[30] D. Aiqattan, "Towards Sign Language Recognition Using EEG-Based Motor Imagery Brain Computer Interface," 2017.

[31] Y. Xu, M. Li, L. Cui, S. Huang, F. Wei, and M. Zhou, "LayoutLM: Pre-training of Text and Layout for Document Image Understanding," in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Association for Computing Machinery, Aug. 2020, pp. 1192–1200. doi: 10.1145/3394486.3403172.

[32] S. V. Mahadevkar, S. Patil, K. Kotecha, L. W. Soong, and T. Choudhury, "Exploring AI-driven approaches for unstructured document analysis and future horizons," *J Big Data*, vol. 11, no. 1, Dec. 2024, doi: 10.1186/s40537-024-00948-z.

[33] J. Zhou, C. Yang, Y. Zhu, and Y. Zhang, "Cross-region feature fusion with geometrical relationship for OCR-based image captioning," *Neurocomputing*, vol. 601, p. 128197, Oct. 2024, doi: 10.1016/J.NEUCOM.2024.128197.

[34] J. Zhou, C. Yang, Y. Zhu, and Y. Zhang, "Cross-region feature fusion with geometrical relationship for OCR-based image captioning," *Neurocomputing*, vol. 601, p. 128197, Oct. 2024, doi: 10.1016/J.NEUCOM.2024.128197.

[35] M. S. Ghafourian, S. Tarkiani, M. Ghajar, M. Chavooshi, H. Khormaei, and A. Ramezani, "Optimizing warfarin dosing in diabetic patients through BERT model and machine learning techniques," *Comput Biol Med*, vol. 186, p. 109755, Mar. 2025, doi: 10.1016/J.COMPBIOMED.2025.109755.

[36] K. I. Roumeliotis, N. D. Tselikas, and D. K. Nasiopoulos, "Optimizing Airline Review Sentiment Analysis: A Comparative Analysis of LLaMA and BERT Models through Fine-Tuning and Few-Shot Learning," *Computers, Materials and Continua*, vol. 82, no. 2, pp. 2769–2792, Feb. 2025, doi: 10.32604/CMC.2025.059567.