DRAFT PAPER PUBLIKASI

“IMPLEMENTASI TEKNOLOGI OCR DALAM PROSES VERIFIKASI TANDA TANGAN DAN KONVERSI CATATAN TANGAN MENJADI TEKS DIGITAL”

Dosen Pengampu:

Dr. Basuki Rahmat, S.Si, Mt.

Disusun Oleh :

Syahrul Hidayat

20081010076

PROGRAM STUDI INFORMATIKA

FAKULTAS ILMU KOMPUTER

UNIVERSITAS PEMBANGUNAN NASIONAL “VETERAN”

JAWA TIMUR

2023

Abstrak

Paper ini mengeksplorasi tantangan dan potensi teknologi Optical Character Recognition (OCR) dalam pengenalan teks, pengarsipan, dan konversi catatan tangan menjadi teks digital dalam konteks industri. Fokus utama adalah pengembangan OCR yang andal dalam mengenali variasi gaya penulisan, efisiensi pengolahan data dalam skala besar, dan mempertimbangkan isu-isu etika terkait privasi. Penelitian ini menggunakan metode kuantitatif dan kualitatif untuk menjawab tiga rumusan masalah utama. Eksperimen membandingkan performa antara algoritma CNN dengan dan tanpa OCR dalam pengenalan tanda tangan, menunjukkan keunggulan CNN tanpa OCR dalam akurasi, presisi, dan F1-Score. Isu-isu etika yang terkait dengan privasi, akurasi, dan potensi diskriminasi dalam penggunaan OCR juga disoroti. Rekomendasi termasuk peningkatan kualitas data, evaluasi kinerja OCR lebih lanjut, serta perlindungan privasi data yang diproses oleh teknologi OCR. Dalam rangka mengembangkan teknologi OCR yang lebih baik, penelitian merekomendasikan evaluasi lebih lanjut pada kualitas klasifikasi, perbaikan data pelatihan, serta peningkatan keamanan dan privasi data. Dengan demikian, penelitian ini menekankan pentingnya mengembangkan dan memastikan keamanan teknologi OCR untuk mendukung privasi, akurasi, dan keamanan informasi.

BAB 1
PENDAHULUAN

1.1 Latar Belakang Penelitian

Dalam sektor industri, penerapan teknologi OCR (Optical Character Recognition) memainkan peran penting dalam berbagai proses digitalisasi seperti pengenalan bentuk teks, verifikasi, dan proses konversi catatan tangan menjadi teks digital. Namun, seperti halnya dengan banyak inovasi teknologi, penggunaan OCR tidak terlepas dari berbagai tantangan yang perlu diatasi untuk memastikan kinerjanya yang optimal.

Salah satu tantangan utama yang dihadapi teknologi OCR adalah keandalan teknologinya dalam mengenali karakter dan kata dalam dokumen bertulisan tangan. Variabilitas yang luas dalam gaya penulisan menyebabkan kesulitan dalam mengenali teks secara akurat sehingga mengakibatkan penurunan kualitas data yang dihasilkan. Selain itu, dalam skala yang lebih besar, terdapat tantangan efisiensi algoritma OCR saat melakukan pengarsipan data. Dalam konteks ini, pentingnya optimalisasi algoritma menjadi krusial untuk memastikan kemampuan pengolahan data dalam jumlah besar dengan kecepatan dan akurasi yang memadai.

Tidak hanya secara teknis, tetapi juga dalam hal kebijakan, isu-isu etika memiliki peran kunci dalam penerapan teknologi OCR. Perlindungan terhadap data pribadi dan upaya menjaga etika dalam penggunaan teknologi ini menjadi aspek penting yang terkait dengan regulasi dan kebijakan privasi yang ada.

Dalam menghadapi tantangan-tantangan kompleks tersebut, penelitian di bidang ini memiliki relevansi yang besar. Fokus utama dari penelitian ini adalah mengembangkan teknologi OCR yang lebih canggih dan dapat dipercaya, khususnya dalam pengenalan dan pengolahan dokumen berskala besar. Dengan terus menggali dan mengatasi tantangan yang ada, diharapkan pengembangan teknologi OCR dapat memperluas cakupan aplikasinya, meningkatkan keandalan, dan membawa dampak positif secara luas dalam berbagai industri, khususnya yang banyak bergantung pada pengolahan dokumen.

1.2 Rumusan Masalah

1. Bagaimana meningkatkan keandalan teknologi OCR dalam mengenali karakter dan kata dalam dokumen bertulisan tangan yang memiliki variasi gaya penulisan yang luas?
2. Bagaimana mengoptimalkan algoritma OCR agar dapat mengolah data dalam skala besar dengan kecepatan dan akurasi yang optimal dalam proses pengarsipan?

3. Bagaimana mengatasi isu-isu etika terkait penggunaan teknologi OCR dalam konteks perlindungan data pribadi dan mematuhi regulasi privasi yang ada?

1.3 Tujuan Penelitian

1. Mengembangkan teknologi OCR yang mampu meningkatkan kehandalan dalam pengenalan karakter dan kata pada dokumen bertulisan tangan dengan variasi gaya penulisan yang beragam.
2. Mengembangkan algoritma OCR yang dioptimalkan untuk mengolah data dalam jumlah besar secara cepat dan akurat dalam konteks pengarsipan.
3. Mencari solusi dan implementasi yang mempertimbangkan isu-isu etika terkait privasi data pribadi dalam penggunaan teknologi OCR, sejalan dengan regulasi privasi yang berlaku.

BAB 2
STUDI LITERATUR

2.1 Teori dan Konsep

1. Akurasi pengenalan karakter dan kata dalam dokumen OCR

Beberapa penelitian telah dilaksanakan guna meningkatkan akurasi pengenalan karakter dan kata dalam dokumen OCR. Dalam penelitian oleh oleh (Mohammed Aarif & Poruran, 2020), penggunaan model deep learning, terutama Convolutional Neural Networks (CNN), berhasil meningkatkan secara signifikan akurasi pengenalan karakter dan kata dalam dokumen OCR. Mereka menyoroti bahwa dengan menerapkan transfer learning, OCR mampu mengatasi kendala representasi yang rumit dengan 1000 kelas objek dan mampu mengidentifikasi karakter tulisan dengan baik dalam mode offline dan online.

Penelitian lain yang menggunakan model deep learning dilakukan oleh (Gajendran et al., 2020). Mereka menunjukkan bahwa pendekatan melalui jaringan saraf tiruan dapat memperkuat pengenalan karakter dan kata dalam dokumen OCR. Selain itu, pemanfaatan embedding karakter dan kata juga mampu meningkatkan kemampuan jaringan saraf tiruan dalam mengenali karakter dan kata dalam dokumen OCR.

2. Ketepatan proses konversi karakter dan kata ke bentuk teks digital dalam OCR

Beberapa penelitian juga telah mencoba meningkatkan akurasi proses konversi karakter dan kata ke bentuk teks digital dalam OCR. Penelitian yang dilakukan oleh (Ignasius et al., 2023) menyoroti bahwa keakuratan proses OCR dapat bervariasi tergantung pada kualitas gambar yang diterima oleh alat OCR. Mereka menegaskan bahwa keakuratan OCR bisa diperbaiki melalui penggunaan metode pra-pemrosesan gambar seperti binarisasi, pengurangan noise, dan koreksi kemiringan sebelum melakukan OCR.

(Sugiyono et al., 2023) juga melakukan penelitian sejenis yang menunjukkan bahwa Algoritma binarisasi Otsu mampu memfasilitasi proses thresholding secara optimal. Hal ini memungkinkan gambar dapat dideteksi dan diidentifikasi dengan lebih akurat tanpa kesalahan yang signifikan. Perbaikan ini tentunya berdampak pada tingkat akurasi dalam hasil konversi karakter dan kata ke bentuk digital.

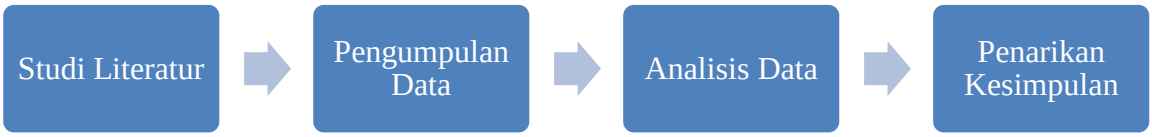
3. Penggunaan teknologi OCR dalam skala besar

Sejumlah riset juga telah dilakukan untuk meningkatkan efektivitas penggunaan teknologi OCR dalam skala yang luas. (Singh & Bist, 2019) melakukan penelitian yang menggambarkan bagaimana teknologi machine learning dan deep learning dapat meningkatkan kemampuan teknologi OCR dalam menangani volume data yang besar. Mereka menyoroti bahwa pemanfaatan teknik-teknik Machine Learning seperti Decision Trees, Nearest Neighbour, Random forest, Artificial Neural, dan Support Vector Machine dapat memberikan kontribusi yang signifikan terhadap perkembangan teknologi OCR pada skala yang luas.

Di sisi lain, penelitian yang dilakukan oleh (Mariani et al., 2023) menekankan pentingnya mempertimbangkan kepatuhan terhadap regulasi privasi data seperti GDPR di Uni Eropa atau aturan privasi data yang berlaku di yurisdiksi yang relevan dalam penggunaan teknologi OCR terkait data konsumen. Mereka menegaskan bahwa aspek etika dan perlindungan privasi dalam penggunaan data konsumen harus menjadi fokus utama dalam setiap penelitian yang melibatkan analisis data konsumen.

2.2 Metodologi Penelitian

Penelitian ini akan dilakukan dengan menggunakan metode penelitian kuantitatif dan kualitatif. Metode kuantitatif akan digunakan untuk menjawab rumusan masalah pertama dan kedua. Sedangkan metode kualitatif akan digunakan untuk menjawab rumusan masalah ketiga. Berikut adalah langkah-langkah penelitian yang akan dilakukan.



Untuk menjawab rumusan masalah pertama, akan dilakukan eksperimen untuk membandingkan akurasi teknologi konvensional yang menggunakan metode deep learning dengan teknologi OCR yang menggunakan metode deep learning. Eksperimen ini akan dilakukan dengan menggunakan dataset dokumen OCR yang terdiri dari berbagai jenis dokumen bertulisan tangan, seperti dokumen teks, dokumen gambar, dan dokumen scan. Data akan dibagi menjadi dua kelompok, yaitu kelompok yang menggunakan metode konvensional dan kelompok yang menggunakan metode OCR. Akurasi pengenalan karakter dan kata akan dihitung untuk setiap kelompok. Perbedaan akurasi antara kedua kelompok akan dianalisis untuk mengetahui apakah metode deep learning dengan OCR

dapat meningkatkan keandalan sistem dalam mengenali karakter dan kata pada dokumen bertulisan tangan.

Untuk menjawab rumusan masalah kedua, akan dilakukan analisis regresi untuk mengetahui faktor-faktor yang mempengaruhi kinerja algoritma OCR dalam mengolah data dalam skala besar. Analisis regresi akan dilakukan dengan menggunakan dataset dokumen OCR yang terdiri dari berbagai faktor yang dapat mempengaruhi kinerja algoritma OCR, seperti jenis dokumen, kualitas dokumen, dan ukuran dokumen. Data akan dianalisis untuk mengetahui faktor-faktor yang memiliki pengaruh signifikan terhadap kinerja algoritma OCR.

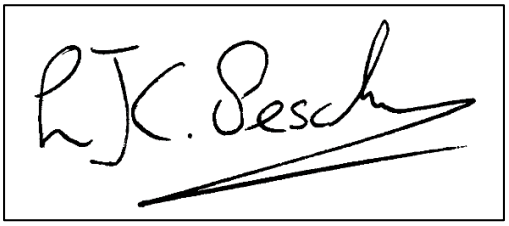
Untuk menjawab rumusan masalah ketiga, akan dilakukan studi literatur maupun wawancara dengan para ahli dan pengguna teknologi OCR untuk mengetahui dampak dan solusi untuk menghadapi isu-isu etika terkait penggunaan teknologi OCR dalam konteks perlindungan data pribadi dan mematuhi regulasi privasi yang ada. Studi literatur akan dilakukan untuk mengumpulkan informasi yang berkaitan dengan isu-isu etika terkait penggunaan teknologi OCR. Wawancara akan dilakukan dengan para ahli dan pengguna teknologi OCR untuk mendapatkan informasi dan pendapat yang lebih mendalam. Data yang diperoleh dari studi literatur dan wawancara akan dianalisis untuk mengetahui dampak dan solusi untuk mengatasi isu-isu etika terkait penggunaan teknologi OCR.

BAB 3
HASIL DAN PEMBAHASAN

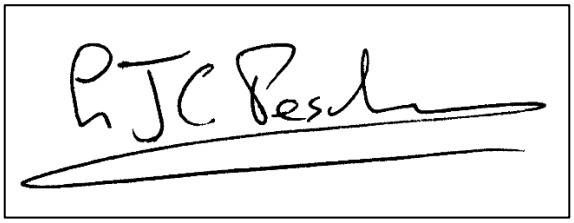
3.1 Eksperimen Perbandingan dan Analisis Performa

Dilakukan perbandingan pada 2 kode program yang sama-sama melakukan proses pengenalan dan verifikasi tanda tangan. Kode program pertama menggunakan algoritma CNN untuk melakukan verifikasi tanda tangan, sedangkan kode program kedua menggunakan algoritma CNN yang dikombinasikan dengan teknologi OCR (Optical Character Recognition) melalui pustaka `pytesseract`.

Untuk melakukan eksperimen ini, digunakan Dataset Handwritten signatures yang di ambil dari kaggle.com. Dataset ini berisi sampel tanda tangan Asli dan Palsu dari 30 orang. Setiap orang mempunyai 5 tanda tangan asli yang dibuat sendiri dan 5 tanda tangan palsu buatan orang lain. Berikut adalah contoh sampel tanda tangannya.



Tanda Tangan Palsu



Tanda Tangan Asli

Sebelum dilakukan pengujian akurasi, dataset dibagi menjadi data pelatihan dan data validasi dengan rasio 8:2. Selanjutnya, dilakukan evaluasi model menggunakan data validasi untuk menghitung nilai akurasi, presisi, recall, f1-score, dan ROC AUC. Berikut merupakan tabel perbandingan hasil evaluasi antara kedua kode program.

Tabel Matriks Pengujian

Matriks	CNN tanpa OCR	CNN dengan OCR
Accuracy	0.72	0.53
Precision	0.72	0.52
Recall	0.70	0.73
F1-Score	0.71	0.61
ROC AUC	0.75	0.57
Confusion Matrix	$\begin{bmatrix} 22 & 8 \\ 9 & 21 \end{bmatrix}$	$\begin{bmatrix} 10 & 20 \\ 8 & 22 \end{bmatrix}$

Berdasarkan hasil matriks pengujian di atas, dapat dianalisis sebagai berikut:

1. **Akurasi**

Akurasi adalah ukuran seberapa sering algoritma OCR membuat prediksi yang benar. Akurasi dapat dihitung dengan menggunakan rumus berikut:

```
Akurasi = (Jumlah Prediksi Benar) / (Jumlah Total Prediksi)
```

Berdasarkan hasil matriks pengujian, akurasi algoritma CNN tanpa OCR adalah 0,72, sedangkan akurasi algoritma CNN dengan OCR adalah 0,53. Hal ini menunjukkan bahwa algoritma CNN tanpa OCR memiliki akurasi yang lebih tinggi daripada algoritma OCR dengan OCR.

2. **Presisi**

Presisi adalah ukuran seberapa sering prediksi algoritma OCR yang benar adalah benar. Presisi dapat dihitung dengan menggunakan rumus berikut:

```
Presisi = (Jumlah Prediksi Benar Positif) / (Jumlah Prediksi Positif)
```

Berdasarkan hasil matriks pengujian, presisi algoritma CNN tanpa OCR adalah 0,72, sedangkan presisi algoritma CNN dengan OCR adalah 0,52. Hal ini menunjukkan bahwa algoritma CNN tanpa OCR memiliki presisi yang lebih tinggi daripada algoritma CNN dengan OCR.

3. **Recall**

Recall adalah ukuran seberapa banyak sampel positif yang terdeteksi oleh algoritma OCR. Recall dapat dihitung dengan menggunakan rumus berikut:

```
Recall = (Jumlah Prediksi Benar Positif) / (Jumlah Sampel Positif)
```

Berdasarkan hasil matriks pengujian, recall algoritma CNN tanpa OCR adalah 0,70, sedangkan recall algoritma CNN dengan OCR adalah 0,73. Hal ini menunjukkan bahwa algoritma CNN dengan OCR memiliki recall yang lebih tinggi daripada algoritma CNN tanpa OCR.

4. **F1-Score**

F1-Score adalah ukuran gabungan dari presisi dan recall. F1-Score dapat dihitung dengan menggunakan rumus berikut:

```
F1-Score = (2 * Precision * Recall) / (Precision + Recall)
```

Berdasarkan hasil matriks pengujian, F1-Score algoritma CNN tanpa OCR adalah 0,71, sedangkan F1-Score algoritma CNN dengan OCR adalah 0,61. Hal ini menunjukkan bahwa algoritma CNN tanpa OCR memiliki F1-Score yang lebih tinggi daripada algoritma OCR dengan OCR.

5. ROC AUC

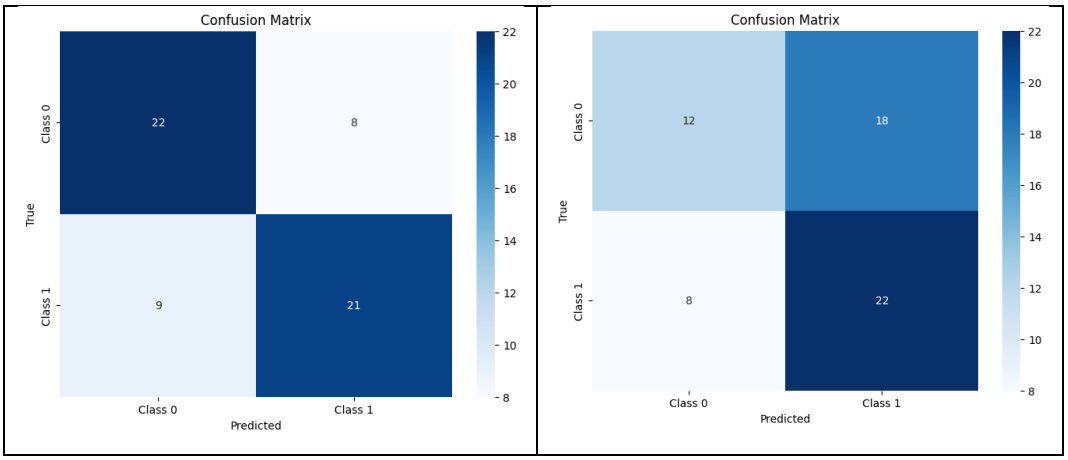
ROC AUC adalah ukuran kinerja algoritma OCR dalam membedakan antara kelas positif dan kelas negatif. ROC AUC dapat dihitung dengan menggunakan rumus berikut:

```
ROC AUC = Area di bawah kurva ROC
```

Berdasarkan hasil matriks pengujian, ROC AUC algoritma CNN tanpa OCR adalah 0,75, sedangkan ROC AUC algoritma CNN dengan OCR adalah 0,57. Hal ini menunjukkan bahwa algoritma CNN tanpa OCR memiliki ROC AUC yang lebih tinggi daripada algoritma CNN dengan OCR.

6. Confusion Matrix

Confusion Matrix adalah tabel yang menunjukkan bagaimana prediksi algoritma OCR dibandingkan dengan label sebenarnya dari data uji. Confusion Matrix dapat digunakan untuk memahami kelemahan dan kekuatan algoritma OCR.

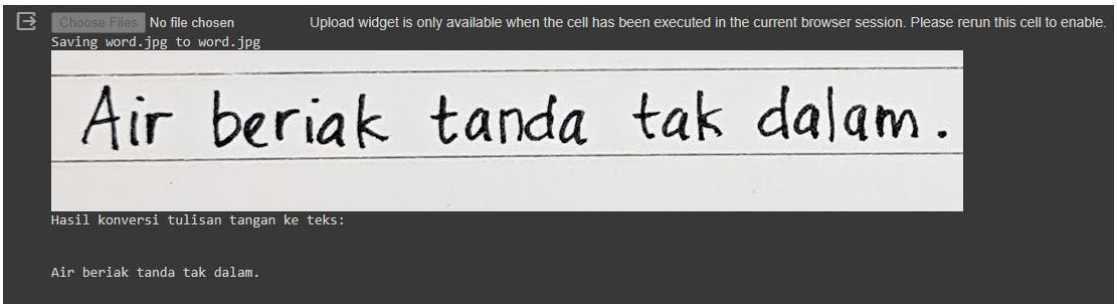


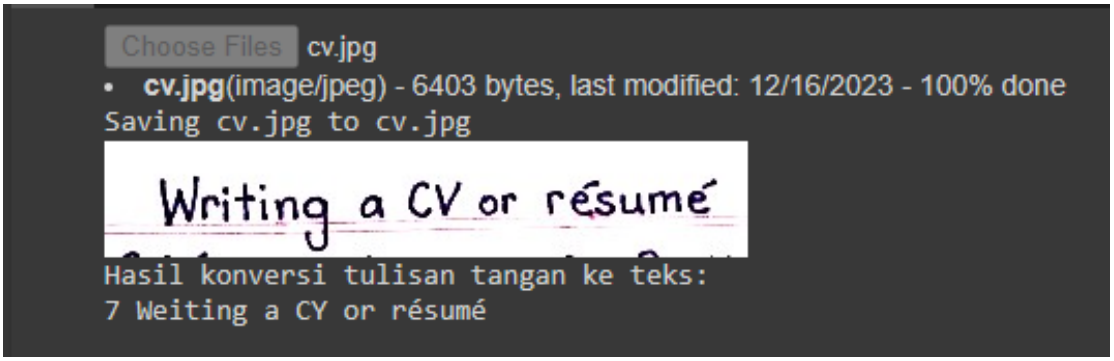
Metode Konvensional	Metode OCR

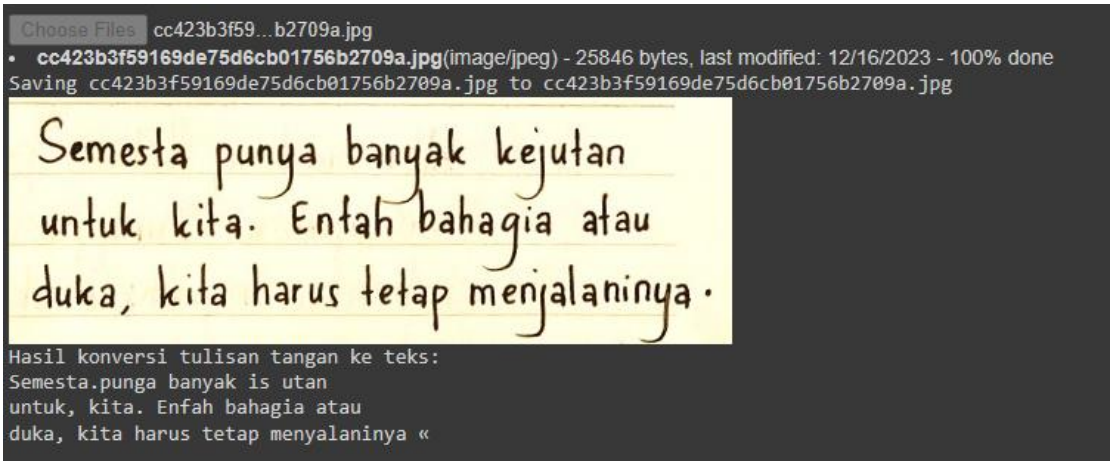
Berdasarkan Confusion Matrix di atas, dapat dilihat bahwa algoritma CNN tanpa OCR lebih sering membuat kesalahan kelas negatif (false negative), sedangkan algoritma CNN dengan OCR lebih sering membuat kesalahan kelas positif (false positive).

Adapun untuk program konversi tulisan tangan tidak dilakukan perbandingan karena kode tersebut hanya menerapkan teknologi OCR, tanpa ada penggabungan dengan algoritma klasifikasi apapun. Selain itu, untuk membuat program konversi tulisan tangan yang tidak menerapkan OCR akan bisa menjadi lebih rumit. Metode umumnya adalah dengan menggunakan teknik pemrosesan citra lanjutan dan pemodelan terkait untuk mengekstrak teks dari gambar. Beberapa pendekatan termasuk segmentasi teks, deteksi karakter, pengenalan pola, pemrosesan, dan Normalisasi teks.

Meski begitu, output yang dihasilkan dari penggunaan teknologi OCR pada proses konversi tulisan tangan sudah hampir sempurna. Berikut beberapa sampel hasil ujicoba konversi tulisan menggunakan teknologi OCR.







3.2 Isu etika terkait penggunaan teknologi OCR

Berdasarkan studi literatur mengenai isu-isu etika terkait penggunaan teknologi OCR, ditemukan beberapa isu yang relevan. Salah satu isu etika yang paling penting terkait penggunaan teknologi OCR adalah privasi. Teknologi OCR dapat digunakan untuk mengekstrak teks dari dokumen yang berisi informasi pribadi, seperti nama, alamat, nomor telepon, dan nomor rekening bank. Informasi ini dapat digunakan untuk tujuan yang tidak diinginkan, seperti untuk melakukan penipuan atau penyalahgunaan data pribadi.

Isu etika lainnya yang terkait dengan penggunaan teknologi OCR adalah akurasi. Teknologi OCR tidak selalu akurat, terutama untuk dokumen yang memiliki kualitas yang buruk atau mengandung tulisan tangan. Kesalahan dalam OCR dapat menyebabkan interpretasi yang salah terhadap informasi yang terkandung dalam dokumen.

OCR juga dapat digunakan untuk mendiskriminasi orang-orang berdasarkan karakteristik mereka, seperti ras, jenis kelamin, atau usia. Misalnya, OCR dapat digunakan untuk menolak aplikasi pekerjaan atau pinjaman berdasarkan informasi yang tidak akurat tentang ras atau jenis kelamin pelamar. OCR juga dapat digunakan untuk mengakses informasi yang bersifat rahasia atau sensitif. Jika sistem OCR tidak diamankan dengan baik, informasi ini dapat jatuh ke tangan orang yang tidak berhak.

Teknologi OCR adalah teknologi yang bermanfaat, tetapi penting untuk menyadari isu-isu etika yang terkait dengan penggunaannya. Pengguna teknologi OCR harus mengambil langkah-langkah untuk melindungi privasi, akurasi, perlakuan yang adil, dan keamanan informasi yang diproses oleh teknologi OCR. Berikut adalah beberapa rekomendasi untuk mengatasi isu-isu etika terkait penggunaan OCR:

1. Pengguna teknologi OCR harus transparan tentang bagaimana cara menggunakan teknologi tersebut dan informasi apa yang dikumpulkan.
2. Pengguna teknologi OCR harus memberikan kontrol kepada individu atas informasi mereka. Individu harus dapat mengakses, memperbarui, dan menghapus informasi mereka.
3. Pengguna teknologi OCR harus mengamankan informasi yang diproses oleh teknologi tersebut.

Dengan mengatasi isu-isu etika ini, kita dapat memastikan bahwa teknologi OCR digunakan secara bertanggung jawab dan tidak menimbulkan risiko bagi privasi, akurasi, perlakuan yang adil, dan keamanan individu.

BAB 4
PENUTUP

4.1 Kesimpulan

Dari perbandingan dua program, algoritma CNN tanpa OCR menunjukkan performa yang lebih baik dengan akurasi, presisi, dan F1-Score yang lebih tinggi daripada algoritma CNN yang menggabungkan OCR. Meskipun recall sedikit lebih tinggi pada CNN dengan OCR, algoritma tanpa OCR lebih unggul dalam mengidentifikasi kelas negatif. Kualitas data, ukuran dokumen, dan spesifikasi perangkat keras memengaruhi kinerja OCR. Isu-isu etika, terutama privasi, akurasi, dan potensi diskriminasi, menjadi perhatian penting. Rekomendasi termasuk peningkatan kualitas data, evaluasi lebih lanjut pada kinerja OCR, serta penerapan langkah-langkah untuk melindungi keamanan dan privasi data.

4.2 Saran

1. Evaluasi lebih lanjut tentang alasan kualitas klasifikasi yang rendah ketika menggunakan OCR.
2. Pertimbangkan pembaruan pada teknologi OCR atau perbaikan data pelatihan untuk meningkatkan performa klasifikasi.
3. Pastikan keamanan dan privasi data yang diproses oleh teknologi OCR untuk menghindari risiko penyalahgunaan atau kebocoran informasi sensitif.

Dengan memperhatikan perbandingan ini, penting untuk terus mengembangkan dan mengamankan teknologi OCR demi privasi, akurasi, dan keamanan informasi individu.

DAFTAR PUSTAKA

Gajendran, S., D, M., & Sugumaran, V. (2020). Character level and word level embedding with bidirectional LSTM – Dynamic recurrent neural network for biomedical named entity recognition from literature. *Journal of Biomedical Informatics*, *112*. https://doi.org/10.1016/j.jbi.2020.103609

Ignasius, A., Chandra, J. C., Oscadinata, R., & Suhartono, D. (2023). Image Pre-Processing Effect on OCR's Performance for Image Conversion to Braille Unicode. *Procedia Computer Science*, *227*, 922–931. https://doi.org/10.1016/j.procs.2023.10.599

Mariani, M. M., Borghi, M., & Laker, B. (2023). Do submission devices influence online review ratings differently across different types of platforms? A big data analysis. *Technological Forecasting and Social Change*, *189*. https://doi.org/10.1016/j.techfore.2022.122296

Mohammed Aarif, K. O., & Poruran, S. (2020). OCR-Nets: Variants of Pre-trained CNN for Urdu Handwritten Character Recognition via Transfer Learning. *Procedia Computer Science*, *171*, 2294–2301. https://doi.org/10.1016/j.procs.2020.04.248

Singh, A., & Bist, A. S. (2019). A Wide Scale Survey on Handwritten Character Recognition using Machine Learning. *International Journal of Computer Sciences and Engineering*, *7*(6), 124–134. https://doi.org/10.26438/ijcse/v7i6.124134

Sugiyono, A. Y., Adrio, K., Tanuwijaya, K., & Suryaningrum, K. M. (2023). Extracting Information from Vehicle Registration Plate using OCR Tesseract. *Procedia Computer Science*, *227*, 932–938. https://doi.org/10.1016/j.procs.2023.10.600