JURNAL RESISTOR | ISSN 2598-7542 | E-ISSN 2598-9650
Vol. 6 No 1 – April 2023 | https://s.id/jurnalresistor
DOI : https://doi.org/10.31598
Publishing : Prahasta Publisher

Implementasi Hirarki Dataset Dalam Membangun Model Language Aksara Bali Menggunakan Framework Tesseract OCR

Ahmad Asroni[1], Gede Indrawan[2], Luh Joni Erawati Dewi[3]

[123]Program Studi Ilmu Komputer, Fakultas Pascasarjana, Universitas Pendidikan Ganesha
Jl. Udayana 11, Singaraja, Buleleng, Bali, Indonesia

e-mail: ahmad.asroni@undiksha.ac.id[1], gindrawan@undiksha.ac.id[2], joni.ernawati@undiksha.ac.id[3]

Received : April, 2023	Accepted : April, 2023	Published : April, 2023

Abstract

The current decline in the use of Balinese Script is due to the lack of interest of the Balinese people in learning it, as the process of recognizing Balinese Script is relatively complicated. Therefore, Optical Character Recognition (OCR) technology has been developed to help overcome this problem. This research aims to implement one of the popular OCR engines, namely Tesseract OCR to recognize Balinese characters. The experimentation process consists of four stages, namely dataset preparation, dataset generation, dataset training, and implementing the language model into a mobile-based platform. The results show that the use of Web Scraping method for dataset collection is superior compared to manual dataset collection. The best language model result is a combination of character, word, sentence, and paragraph datasets with an accuracy rate of 70.37%. The accuracy rate will be higher if the dataset is more diverse and the hierarchy level is more structured. This research can provide a solution to the problem of decreasing the use of Balinese characters by utilizing OCR technology to facilitate the process of recognizing Balinese characters. In addition, the results of this study can also be used as a reference for the development of better OCR technology in the future.

Keywords: *aksara bali, optical character recognition, tesseract ocr, web scraping, dataset hierarchy*

Abstrak

Penurunan penggunaan Aksara Bali saat ini disebabkan oleh kurangnya minat masyarakat Bali dalam mempelajarinya, karena proses pengenalan Aksara Bali relatif rumit. Oleh karena itu, teknologi Optical Character Recognition (OCR) telah dikembangkan untuk membantu mengatasi masalah ini. Penelitian ini bertujuan untuk mengimplementasikan salah satu mesin OCR populer, yaitu Tesseract OCR untuk mengenali karakter Aksara Bali. Proses ekperimen terdiri dari empat tahap, yaitu persiapan dataset, generate dataset, training dataset, dan mengimplementasikan language model ke dalam platform berbasis mobile. Hasil penelitian menunjukkan bahwa penggunaan metode Web Scraping untuk pengumpulan dataset lebih unggul dibandingkan dengan pengumpulan dataset secara manual. Hasil model language terbaik yang dihasilkan adalah kombinasi dataset karakter, kata, kalimat, dan paragraf dengan tingkat akurasi sebesar 70,37%. Tingkat akurasi akan semakin tinggi jika dataset semakin beragam dan tingkat hirarkinya semakin terstruktur. Penelitian ini dapat memberikan solusi untuk mengatasi masalah penurunan penggunaan Aksara Bali dengan memanfaatkan teknologi OCR untuk memudahkan proses pengenalan karakter Aksara Bali. Selain itu, hasil penelitian ini juga dapat digunakan sebagai acuan untuk pengembangan teknologi OCR yang lebih baik di masa depan.

Kata Kunci: *aksara bali, optical character recognition, tesseract ocr, web scraping, hirarki dataset*

1. PENDAHULUAN

Aksara Bali adalah salah satu bentuk tulisan kuno yang masih digunakan di Bali dan beberapa daerah lainnya di Indonesia. Aksara Bali terdiri dari 47 karakter dasar (aksara swara dan aksara konsonan) dan 18 tanda baca. Karakter aksara Bali ditulis dari atas ke bawah dan dari kiri ke kanan, dan biasanya digunakan untuk menulis bahasa Bali, serta bahasa Sanskerta, Kawi, dan Jawa Kuno. Meskipun memiliki nilai sejarah yang tinggi, namun penggunaan aksara Bali seringkali mengalami kendala dalam era digital saat ini. Beberapa masalah yang sering dihadapi adalah kompleksitas dalam membaca dan menulis aksara Bali, keterbatasan dalam pembelajaran dan penggunaan teknologi, serta kekhawatiran akan kepunahan aksara Bali [1]. Namun demikian, upaya pelestarian aksara Bali terus dilakukan oleh beberapa pihak. Pemerintah Bali telah memasukkan pengajaran aksara Bali sebagai mata pelajaran wajib di sekolah-sekolah di Bali. Selain itu, beberapa teknologi seperti aplikasi penerjemah bahasa Bali, font aksara Bali untuk komputer dan smartphone, serta pengenalan teks aksara Bali menggunakan teknologi deep learning juga telah dikembangkan untuk membantu dalam penggunaan aksara Bali di era digital [2].

Pengenalan karakter optik, juga dikenal sebagai Optical Character Recognition (OCR), telah menjadi semakin lazim dengan perkembangan teknologi komputer. OCR adalah teknik untuk mengubah teks dan gambar yang dicetak menjadi karakter digital yang dapat dimanipulasi oleh mesin. Berbagai sektor aplikasi, seperti pendidikan, perbankan, keuangan, dan hukum, telah menerapkan OCR. Karena aksara Latin Inggris didukung oleh standar American Standard Code for Information Interchange Encoding, atau disingkat ASCII, sebagian besar pengembangan OCR terus berkonsentrasi pada aksara tersebut. Keterbatasan kemampuan OCR untuk mengenali aksara non-Latin menyulitkan para peneliti untuk berimprovisasi. Seiring dengan perkembangan teknologi OCR, banyak penelitian yang memanfaatkan OCR untuk pengenalan karakter aksara non-Latin [3].

Dalam penelitiannya, Abdul Robby dkk mengimplementasikan Tesseract OCR sebagai mesin pengenal aksara Jawa. Penelitian ini berusaha untuk mempermudah proses pengenalan aksara Jawa secara otonom melalui aplikasi mobile [4]. Dataset data latih (traineddata) yang digunakan untuk membangun mesin Tesseract OCR berisi 5.880 aksara Jawa. Untuk membangun dataset Aksara Jawa, karakter digital dengan spesifikasi (3 set x 120 karakter) dan tulisan tangan (46 set x 120 karakter) dikumpulkan. Penelitian ini menggunakan Neural-Network API dari mesin Tesseract OCR untuk melatih dataset. Dengan menggunakan JTessBoxEditor, dataset alfabet Jawa dipilih sebelum prosedur pelatihan dengan melakukan segmentasi pada setiap karakter dan mengatur variabel untuk setiap kluster karakter. Model yang dihasilkan dari data yang dilatih mencapai tingkat akurasi maksimum sebesar 97,50%.

Penelitian Mudiarta dkk. merupakan penelitian pembanding berikutnya untuk kasus pengenalan karakter optik non-Latin. Menggabungkan teknologi informasi dengan disiplin ilmu aksara Bali, penelitian ini berkonsentrasi pada pelestarian pengetahuan untuk memahami aksara Bali pada gambar. Dalam penelitian ini, aplikasi OCR dikembangkan pada perangkat mobile yang dilengkapi kamera. Aplikasi ini menerima gambar sebagai masukan dan memprosesnya menggunakan teknologi mesin OCR Tesseract. Untuk menjalankan prosedur pelatihan, dataset Aksara Bali dibangun dengan hanya menggunakan angka dan delapan belas suku kata dasar. Prosedur ini dilakukan dengan menggunakan jTessBoxEditor, yang memiliki kemampuan otomatis untuk pelatihan dataset. Dengan font Aksara Bali-Simbar berbasis gambar berkualitas tinggi, 62% dari 50 kata uji dikenali dengan benar [1].

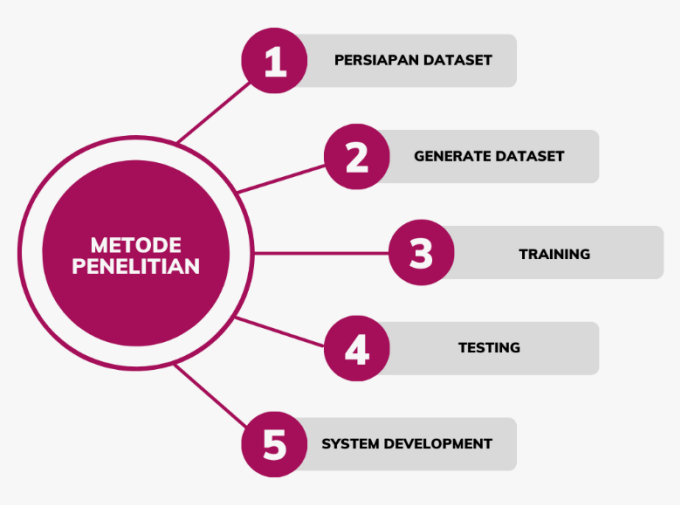
Dari pemaparan dua penelitian diatas terdapat kemiripan dari sisi engine Optical Character Recognition yang digunakan dan proses training data yang dilakukan. Training data yang dilakukan untuk membuat model traineddata memanfaatkan tools jTessBoxEditor dengan cara melakukan segmentasi karakter dari gambar karakter non latin. Proses segmentasi tersebut dilakukan secara bergantian untuk masing-masing dataset yang dimiliki. Ada beberapa kelemahan yang terjadi pada dua penelitian tersebut khususnya pada proses training data yang dilakukan. Penggunaan tools jTessBoxEditor harus dilakukan dengan manual dengan melakukan segmentasi untuk masing-

masing dataset membuat proses training relative menjadi lebih membutuhkan waktu yang lama. Pada bagian bab saran dari dua penelitian tersebut berfokus kepada peningkatan jumlah dataset yang digunakan.

Berdasarkan kelemahan dan saran dari dua penelitian tersebut dapat disolusikan dengan menggunakan metode training data yang berbeda selain menggunakan tools jTessBoxEditor ada metode training terbaru untuk membuat traineddata yaitu dengan menggunakan metode training Tesseract OCR terbaru. Metode training Tesseract OCR terbaru ini dapat melakukan training dataset secara simultan untuk seluruh dataset. Menurut Idrees & Hassani sejak versi 4.0, Tesseract OCR menghadirkan mesin baru berbasis Long Short-Term Memory (LSTM) [5]. LSTM, sebagai bentuk khusus dari Jaringan Syaraf Tiruan (RNN), memberikan akurasi yang jauh lebih tinggi pada pengenalan gambar dari pada versi Tesseract OCR sebelumnya. Tesseract bisa dilatih dari awal atau disempurnakan berdasarkan bahasa yang sudah terlatih.

2. METODE PENELITIAN

Penelitian ini berkonsentrasi pada penerapan model pelatihan Tesseract OCR terbaru untuk karakter digital non-Latin, terutama untuk bahasa yang belum didukung oleh Tesseract OCR dan yang belum pernah diteliti sebelumnya. Penelitian ini menggunakan metode pelatihan data terbaru dari Tesseract OCR, yang berkonsentrasi pada format dataset gambar dan ground truth. Prosedur pelatihan ini berbeda dengan dua penelitian [1] dan [4] yang meneliti pengenalan karakter digital non-Latin dengan menggunakan jTessBoxEditor sebagai alat data pelatihan. Gambar 1 menggambarkan tahapan-tahapan yang dilakukan dalam penelitian ini.



Gambar 1. Metode Penelitian

2.1 Persiapan Dataset

Proses persiapan dataset merupakan tahapan untuk melakukan konversi dari bahasa bali atau latin bali ke aksara bali. Proses ini menghasilkan dataset yang berisi gambar skrip dan ground truth. G. Indrawan dkk. [6] melakukan penelitian yang menjadi sumber data yang digunakan untuk menghasilkan dataset. G. Indrawan dkk. melakukan penelitian yang menghasilkan lebih dari 35.000 kata dataset dalam bahasa Bali, Indonesia, dan Inggris. Prosedur transliterasi dalam penelitian ini diimplementasikan dengan menggunakan platform yang berbeda, khususnya platform berbasis web. Beberapa langkah diperlukan untuk mempersiapkan dataset ini untuk ditransliterasi dari aksara Latin ke aksara Bali. Langkah awal adalah mengonversi dataset Latin Bali yang masih ada di database ke Unicode sehingga dapat ditampilkan pada halaman HTML. Selain itu, menambahkan keluarga Font Bali Noto Sans sehingga unicode yang ditampilkan di halaman HTML dapat dikonversi ke dalam karakter aksara Bali digital [7][8]. Proses berikutnya yang dilakukan setelah konversi dataset adalah melakukan generate dataset menggunakan metode web scraping. Web scraping adalah

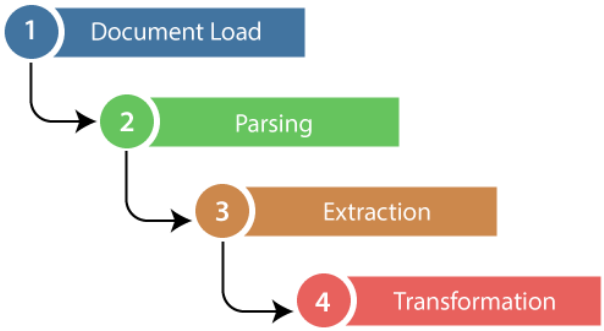
teknik yang umum digunakan untuk menghasilkan dataset yang mengekstrak informasi dari platform web [9]. Teknik web scraping digunakan untuk mengekstrak informasi secara mandiri dari situs web dengan mengurai elemen hypertext dan mengambil informasi dalam bentuk teks, gambar, dan video yang disematkan di dalamnya dari sejumlah besar data pada halaman web [10][11]. Penelitian ini menggunakan teknik web crawling yang terdiri dari empat prosedur utama [12]. Pertama, sebuah halaman HTML yang berisi informasi yang dapat diekstraksi menjadi dataset gambar aksara Bali dan kebenaran dasar aksara Bali dibuat sebagai template scraping [13]. Prosedur kedua melibatkan pengaksesan situs web melalui kolom pencarian pada browser. Proses ketiga mengembangkan algoritma web mining untuk mendapatkan gambar aksara Bali dan secara otomatis mengekstrak kebenaran dasar ketika algoritma dijalankan. Langkah terakhir adalah menyimpan semua dataset yang dihasilkan oleh teknik web scraper ke dalam database. Dua kategori dataset dihasilkan dari proses pemindaian web: dataset gambar aksara Bali dan format karakter digital aksara Bali yang benar. Daftar suku kata dan angka dasar aksara bali dapat dilihat pada Gambar 2, dan Ilustrasi proses web scraping dapat dilihat pada Gambar 3.

Latin	Balinese
ha	ᬳ
na	ᬦ
ca	ᬘ
ra	ᬭ
ka	ᬓ
da	ᬤ
ta	ᬢ
sa	ᬲ
wa	ᬯ

Latin	Balinese
la	ᬮ
ma	ᬫ
ga	ᬕ
ba	ᬩ
nga	ᬗ
pa	ᬧ
ja	ᬚ
ya	ᬬ
nya	ᬜ

Latin	Balinese
0	᭐
1	᭑
2	᭒
3	᭓
4	᭔
5	᭕
6	᭖
7	᭗
8	᭘
9	᭙

Gambar 2. Suku Kata dan Angka Dasar Aksara Bali

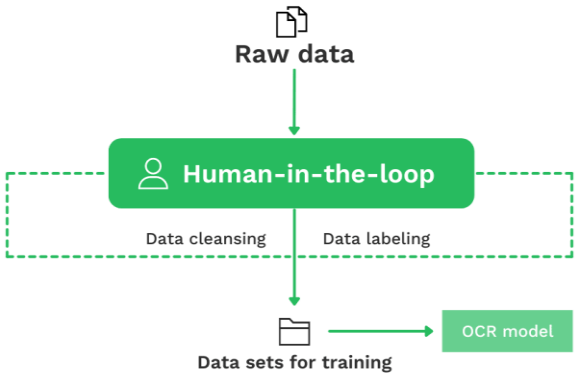


Gambar 3. Proses Web Scraping
[Sumber: https://www.javatpoint.com/web-scraping-using-python]

2.2 Training Dataset

Tesseract adalah sebuah mesin pengenalan karakter optik (OCR) yang dikembangkan oleh Google. Ini digunakan untuk mengenali dan mengekstraksi teks dari gambar atau dokumen yang dipindai. Salah satu komponen penting dalam Tesseract adalah training dataset.

Training dataset adalah proses melatih model pengenalan karakter optik menggunakan kumpulan dataset. Dataset ini terdiri dari gambar-gambar teks yang beragam, termasuk berbagai font, ukuran, dan gaya tulisan. Tujuan dari melatih model menggunakan dataset ini adalah untuk mengajarkan Tesseract cara mengenali dan memahami karakter dalam berbagai kondisi visual. Tesseract OCR dapat dilatih di berbagai sistem operasi, termasuk Linux, Windows, dan macOS, dengan mengeksekusi kumpulan skrip baris perintah dan shell [14]. Tesseract OCR sendiri merekomendasikan penggunaan sistem operasi Linux secara lokal atau di cloud, meskipun ada berbagai pilihan yang tersedia. Virtual server merupakan server dengan performa yang cukup baik ketika melakukan training data. Containers memiliki sejumlah keunggulan yang membuatnya populer di kalangan alat pelatihan data, termasuk konfigurasi yang sederhana, tingkat keamanan yang tinggi, kemampuan untuk berjalan di berbagai platform cloud, kemampuan diagnostik, dan dukungan untuk berbagai sistem operasi [15]. Prosedur pelatihan dataset terdiri dari dua proses utama: pelatihan bentuk karakter dan kompilasi kamus bahasa [16]. Hasil dari pelatihan ini adalah model language Tesseract yang telah dilatih secara khusus untuk mengenali karakter dalam dataset tersebut. Model ini dapat digunakan untuk mengenali teks pada gambar baru yang tidak termasuk dalam dataset pelatihan. Ilustrasi training dataset dapat dilihat pada Gambar 4.



Gambar 4. Proses Training Dataset
[Sumber: https://www.klippa.com/en/blog/information/tesseract-ocr]

2.3 Testing Language Model

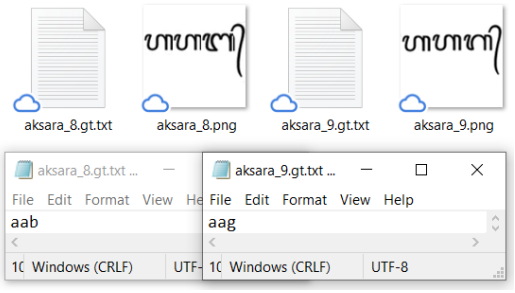
Fase pengujian model language merupakan fase penting yang dilakukan untuk menguji model language yang telah dihasilkan dari fase pelatihan dataset. Hasil data terlatih yang diperoleh setelah melatih dataset menjadi sasaran dari dua kategori pengujian, yaitu pengujian unit dan pengujian performa [17][18]. Persyaratan tambahan diperlukan untuk melakukan pengujian unit otomatis. Ini termasuk ketergantungan tambahan untuk alat pelatihan dan mengunduh semua submodul yang diperlukan, seperti git, selain repositori model. Sementara itu, pengujian kinerja dilakukan untuk menentukan efisiensi dan kemanjuran model berdasarkan alokasi sumber dayanya [14]. Pengujian nilai coincidence, atau biasa disebut sebagai "coincidence testing," adalah suatu pendekatan yang digunakan untuk menguji sejauh mana hasil pengenalan teks dari Tesseract OCR berkorelasi atau cocok dengan teks acuan yang diharapkan. Dalam konteks OCR, pengujian ini bertujuan untuk mengevaluasi akurasi Tesseract dalam mengenali dan mengekstraksi teks dari gambar atau dokumen [19][20]. Hasil pengenalan teks yang diperoleh dari Tesseract kemudian dibandingkan dengan teks acuan yang telah ditentukan sebelumnya. Dalam proses perbandingan ini, metrik evaluasi seperti akurasi karakter, akurasi kata, presisi, recall, atau F1-score dapat digunakan untuk mengukur tingkat kesesuaian antara hasil pengenalan Tesseract dan teks acuan. [21].

3. HASIL DAN PEMBAHASAN

3.1 Ground Truth Dataset

Pada tahap menghasilkan dataset, digunakan platform berbasis web dengan memanfaatkan framework Laravel sebagai backend. Selain

menggunakan backend pada tahap ini juga menggabungkan modul lain untuk proses akuisisi citra yang beroperasi di sisi klien. Penggunaan plugin ini bertujuan untuk meringankan beban server dalam menghasilkan dataset dalam jumlah besar. Proses akuisisi gambar ini menangkap halaman HTML yang dipilih berdasarkan id indeks dari setiap elemen secara bersamaan. Penggunaan id pada setiap elemen HTML bertujuan untuk memberikan identitas yang unik sehingga pada saat modul akuisisi gambar menangkap gambar, modul ini dapat menentukan batas-batas area yang harus diakuisisi. Contoh dataset yang dihasilkan dapat dilihat pada Gambar 5.



Gambar 5. Ground Truth Dataset

3.2 Training Dataset

Pada tahap training dataset, ada beberapa hal yang harus dilakukan pada dataset yang telah dihasilkan. Yang pertama adalah mengelompokkan dataset ke dalam beberapa kategori, yaitu kelompok dataset per karakter, kelompok dataset per kata, kelompok dataset per kalimat dan yang terakhir adalah kelompok dataset per paragraf. Selanjutnya, setelah mengelompokkan dataset, dataset disusun berdasarkan hirarki dataset. Proses penyusunan hirarki dataset dibuat menjadi beberapa versi dengan menguji apakah susunan hirarki tersebut dapat memberikan peningkatan kualitas hasil dataset training. Susunan hirarki dataset training yang pertama adalah susunan hirarki dengan cara mengkombinasikan dataset secara acak (Random Dataset Combination Hierarchy), susunan hirarki dataset ini dilakukan tanpa memperhatikan urutan atau kelompok dataset. Selanjutnya adalah susunan hirarki dataset dengan menggunakan per karakter saja (Single Character Dataset Combination Hierarchy), susunan hirarki dataset ini terdiri dari kombinasi dataset secara acak. Hirarki dataset yang terakhir adalah susunan hirarki yang terdiri dari kelompok dataset per karakter, kelompok dataset per kata, kelompok dataset per kalimat dan yang terakhir adalah kelompok dataset per paragraf (Hirarki Kombinasi Dataset Karakter, Kata, Kalimat dan Paragraf). Konfigurasi hirarki dataset ini memperhatikan urutan level sesuai dengan urutan yang telah dijelaskan sebelumnya. Proses pelatihan dataset dengan menggunakan susunan hirarki ini dilakukan beberapa kali iterasi pelatihan hingga semua level hirarki habis, level pertama yang akan dilatih adalah level dataset per kata, kemudian setelah proses tersebut selesai maka akan dilanjutkan ke level dataset per kata, setelah itu dataset per kalimat, dan yang terakhir adalah level dataset per paragraf. Hasil yang didapatkan dari training dataset menggunakan hirarki ini mengalami peningkatan jika dibandingkan dengan dua ujicoba hirarki sebelumnya. Paradigma bahasa ini akan menjadi language model dari mesin OCR Tesseract. Berdasarkan hasil data training, dapat dilihat bahwa beberapa skenario dataset training dilakukan dengan berbagai komposisi dan hirarki dataset. Hasil model language (traineddata) yang akan digunakan adalah model language yang memiliki tingkat coincidence terbesar.

3.3 Testing Model Language

Faktor utama yang mempengaruhi peningkatan performa coincidence dari ketiga percobaan yang dilakukan dengan menggunakan kombinasi dataset yang berbeda adalah kombinasi dan hirarki dataset yang digunakan. Hasil dari ketiga percobaan tersebut memiliki benang merah dalam hal struktur hirarki dataset, semakin terstruktur hirarki yang digunakan akan memberikan tingkat coincidence yang lebih tinggi. Peningkatan ini disebabkan karena Tesseract OCR mempelajari dan mengenali karakter mulai dari unit terkecil yaitu per karakter, kemudian per kata, setelah itu per

kalimat dan terakhir per paragraf. Penilaian terhadap model language yang dihasilkan berisi tiga skenario pengujian, yaitu mengevaluasi Suku Kata Dasar dan Angka. Dari prosedur pengujian model language tersebut, didapatkan tingkat coincidence untuk skenario pengujian angka tunggal dengan jumlah 10 angka dengan tingkat akurasi sebesar 100%, berikutnya skenario pengujian karakter tunggal dengan jumlah 18 karakter didapatkan tingkat coincidence sebesar 100%, dan tingkat terakhir dengan skenario pengujian kata dengan jumlah 100 kata didapatkan tingkat coincidence sebesar70,37%. Hasil pengujian model language yang dihasilkan dapat dilihat lebih lanjut pada Tabel 1 disajikan dalam bentuk tabel.

Tabel 1: Testing Model Language

Skenario Pengujian	Jumlah	Benar	Salah	Akurasi
Angka Tunggal	10	10	0	100%
Karakter Tunggal	18	18	0	100%
Kata	54	38	16	70,37%

4. KESIMPULAN

Berdasarkan hasil investigasi ini, ada beberapa kesimpulan yang dapat diambil. Pada proses penyiapan dataset, dilakukan beberapa tahap awal yaitu menyiapkan data transliterasi bahasa Bali, kemudian melakukan konversi huruf latin bali ke aksara Bali menggunakan unicode dan yang terakhir adalah mengembangkan template untuk proses generate dataset. Proses generate dataset menggunakan metode web scraper dan platform berbasis website untuk proses akuisisi citra. Hasil dari generate dataset adalah sepasang data set: sinyal gambar teks satu baris dengan ekstensi file .png dan transkripsi satu baris dengan ekstensi file .gt.txt. Jumlah set data yang dihasilkan secara efektif adalah 35.319. Metode dan mesin pengenalan karakter optik yang digunakan dalam pelatihan dataset serta proses pengenalan isyarat adalah Tesseract OCR versi 5. Prosedur pelatihan dataset terdiri dari tiga kali percobaan dengan struktur hirarki dataset yang berbeda-beda.Testing model language dilakukan menggunakan tiga skenario yaitu skenario pengujian angka tunggal, skenario pengujian karakter tunggal dan skenario pengujian kata. Tingkat akurasi yang didapat oleh skenario pengujuan angka tunggal dan karakter tunggal adalah sebesar 100%. Namun mengalami penuruan ketika menggunakan skenario pengujuan kata yang mendapat tingkat akurasi sebesar 70,37%. Hasil dari prosedur pelatihan tersebut diimplementasikan ke dalam platform aplikasi berbasis mobile. Pengembangan aplikasi mobile menggunakan framework mobile Flutter dengan menerapkan konsep arsitektur clean code. Aplikasi mobile memiliki beberapa halaman utama: Layar Kamera, Layar Pratinjau Gambar, Layar Aksara Bali, dan Layar Sejarah. Oleh karena itu, dapat disimpulkan bahwa proses generate dataset dapat menjadi pilihan yang lebih baik jika dihadapkan pada pelatihan data set yang membutuhkan data set yang besar dibandingkan dengan beberapa penelitian terdahulu yang sudah ada yang sudah menunjukkan bahwa menggunakan tools jTessBox relatif membutuhkan waktu yang lebih lama karena harus mengeliminasi data set per karakter.

Berdasarkan hasil dari proses penelitian yang sudah dilakukan, disadari bahwa tingkat coincidence masih jauh dari kata positif. Ada beberapa hal yang menjadi pertimbangan penting untuk dilakukan dalam meningkatkan hasil coincidence. Pada penelitian ini, dataset yang digunakan dalam membuat model language dibatasi hanya menggunakan data gambar sintetis. Pekerjaan selanjutnya yang akan dilakukan adalah membangun beberapa hirarki dataset dengan mengkombinasikan beberapa karakter aksara bali dengan berbagai bentuk baik itu karakter optik, data asli maupun tulisan tangan aksara bali. Hirarki dataset akan mengacu pada norma-norma penulisan.

PERNYATAAN PENGHARGAAN

Dalam penulisan artikel ini, penulis ingin menyampaikan ucapan terima kasih yang tulus kepada pembimbing Gede Indrawan dan Luh

Joni Erawati Dewi yang telah memberikan bimbingan, arahan, dan dukungan selama proses penulisan artikel. Tanpa bimbingan yang diberikan, penulisan artikel ini tidak akan terwujud dengan baik. Terima kasih juga kepada pihak-pihak lain yang turut berperan dalam proses penulisan artikel ini, baik secara langsung maupun tidak langsung. Semua kontribusi mereka sangat berarti bagi kesuksesan penulisan artikel ini.

DAFTAR PUSTAKA

[1] I. M. D. R. Mudiarta *et al.*, "Balinese character recognition on mobile application based on tesseract open source OCR engine," *J. Phys. Conf. Ser.*, vol. 1516, no. 1, 2020, doi: 10.1088/1742-6596/1516/1/012017.

[2] Gubernur Bali, *Peraturan Gubernur Bali Nomor 80*. Indonesia, 2018.

[3] A. Qaroush, A. Awad, M. Modallal, and M. Ziq, "Segmentation-based, omnifont printed Arabic character recognition without font identification," *J. King Saud Univ. - Comput. Inf. Sci.*, no. xxxx, 2020, doi: 10.1016/j.jksuci.2020.10.001.

[4] G. Abdul Robby, A. Tandra, I. Susanto, J. Harefa, and A. Chowanda, "Implementation of optical character recognition using tesseract with the javanese script target in android application," *Procedia Comput. Sci.*, vol. 157, pp. 499–505, 2019, doi: 10.1016/j.procs.2019.09.006.

[5] S. Idrees and H. Hassani, "Exploiting script similarities to compensate for the large amount of data in training tesseract lstm: Towards kurdish ocr," *Appl. Sci.*, vol. 11, no. 20, 2021, doi: 10.3390/app11209752.

[6] G. Indrawan, N. N. H. Puspita, I. K. Paramarta, and Sariyasa, "LBtrans-bot: A Latin-to-Balinese script transliteration robotic system based on noto sans Balinese font," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 12, no. 3, pp. 1247–1256, 2018, doi: 10.11591/ijeecs.v12.i3.pp1247-1256.

[7] I. K. Paramarta, G. Indrawan, I. B. Rai, and I. N. Martha, "Bound Vowels Grapheme Representation in Balinese Script," in *Proceedings of the 2nd International Conference on Languages and Arts across Cultures (ICLAAC 2022)*, Atlantis Press SARL, 2023, pp. 165–172.

[8] G. Indrawan, L. J. E. Dewi, I. Gede Aris Gunadi, K. Agustini, and I. Ketut Paramarta, "The Analysis of Noto Serif Balinese Font to Support Computer-assisted Transliteration to Balinese Script," in *Lecture Notes in Networks and Systems*, 2023, vol. 400, pp. 571–580, doi: 10.1007/978-981-19-0095-2_54.

[9] S. Chaudhari, R. Aparna, V. G. Tekkur, G. L. Pavan, and S. R. Karki, "Ingredient/Recipe Algorithm using Web Mining and Web Scraping for Smart Chef," *Proc. CONECCT 2020 - 6th IEEE Int. Conf. Electron. Comput. Commun. Technol.*, no. 3, pp. 22–25, 2020, doi: 10.1109/CONECCT50063.2020.9198450.

[10] W. Uriawan, A. Wahana, D. Wulandari, W. Darmalaksana, and R. Anwar, "Pearson correlation method and web scraping for analysis of islamic content on instagram videos," *Proc. - 2020 6th Int. Conf. Wirel. Telemat. ICWT 2020*, 2020, doi: 10.1109/ICWT50448.2020.9243626.

[11] G. Adomavicius and A. Tuzhilin, "Web Scraping:State of the art," *IEEE Trans. Knowl. Data Eng.*, vol. 17, no. 6, pp. 734–749, 2019.

[12] I. B. G. W. Antara, G. R. Dantes, I. G. A. Gunadi, and A. A. G. B. Ariana, "Effect of image partitioning on content-based image retrieval using colour and texture," in *Journal of Physics: Conference Series*, Jun. 2020, vol. 1516, no. 1, doi: 10.1088/1742-6596/1516/1/012015.

[13] G. Indrawan, A. Asroni, L. Joni Erawati Dewi, I. G. A. Gunadi, and I. K. Paramarta, "Balinese Script Recognition Using Tesseract Mobile Framework," *Lontar Komput. J. Ilm. Teknol. Inf.*, vol. 13, no. 3, p. 160, Nov. 2022, doi: 10.24843/lkjiti.2022.v13.i03.p03.

[14] C. Clausner, A. Antonacopoulos, and S. Pletschacher, "Efficient and effective OCR engine training," *Int. J. Doc. Anal. Recognit.*, vol. 23, no. 1, pp. 73–88, 2020, doi: 10.1007/s10032-019-00347-8.

[15] V. K. Kaliappan, S. Yu, R. Soundararajan, S. Jeon, D. Min, and E. Choi, "High-

Secured Data Communication for Cloud Enabled Secure Docker Image Sharing Technique Using Blockchain-Based Homomorphic Encryption," *Energies*, vol. 15, no. 15, 2022, doi: 10.3390/en15155544.

[16] B. Y. Panchal and G. Chauhan, "Design and implementation of android application to extract text from images by using tesseract for English and Hindi," *J. Phys. Conf. Ser.*, vol. 1973, no. 1, 2021, doi: 10.1088/1742-6596/1973/1/012008.

[17] N. H. Khan and A. Adnan, "Urdu optical character recognition systems: Present contributions and future directions," *IEEE Access*, vol. 6, pp. 46019–46046, 2018, doi: 10.1109/ACCESS.2018.2865532.

[18] K. O. Mohammed Aarif and S. Poruran, "OCR-Nets: Variants of Pre-trained CNN for Urdu Handwritten Character Recognition via Transfer Learning," *Procedia Comput. Sci.*, vol. 171, no. 2019, pp. 2294–2301, 2020, doi: 10.1016/j.procs.2020.04.248.

[19] B. Wang, Y. W. Ma, and H. T. Hu, "Hybrid model for Chinese character recognition based on Tesseract-OCR," *Int. J. Internet Protoc. Technol.*, vol. 13, no. 2, pp. 102–108, 2020, doi: 10.1504/IJIPT.2020.106316.

[20] R. Bassam *et al.*, "Autonomous Assistance System for Visually Impaired using Tesseract OCR & gTTS Autonomous Assistance System for Visually Impaired using Tesseract OCR & gTTS," 2022, doi: 10.1088/1742-6596/2327/1/012065.

[21] D. Sporici, E. Cus, and C. Boiangiu, "SS symmetry Using Convolution-Based Preprocessing," 2020.