Building of Informatics, Technology and Science (BITS)
Volume 8, No 1, June 2026 Page: 318−329
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v8i1.9917

Analisis Spasio-Temporal Berbasis Data Video untuk Identifikasi Bangunan Melayu Menggunakan Metode Hybrid CNN-LSTM

Ines Triseptiani, Sri Winiarti*

Fakultas Teknologi Industri, Program Studi Informatika, Universitas Ahmad Dahlan, Yogyakarta, Indonesia
Email: [1]2200018403@webmail.uad.ac.id, [2,*]sri.winiarti@tif.uad.ac.id
Email Penulis Korespondensi: sri.winiarti@tif.uad.ac.id
Submitted: **11/05/2026**; Accepted: **22/06/2026**; Published**: 23/06/2026**

Abstrak−Bangunan Melayu memiliki karakteristik arsitektur yang khas dan perlu didukung dengan sistem identifikasi berbasis teknologi untuk membantu proses dokumentasi serta pelestarian budaya. Penelitian ini bertujuan untuk membangun sistem identifikasi bangunan Melayu dan non-Melayu berbasis data video yang diekstraksi menjadi citra *frame-by-frame*. Penggunaan data video dalam penelitian ini tidak dimaksudkan untuk menganalisis pergerakan objek bangunan, melainkan untuk memanfaatkan variasi tampilan visual yang muncul akibat perubahan sudut kamera, jarak pengambilan gambar, pencahayaan, komposisi objek, dan bagian bangunan yang terekam. Metode yang digunakan adalah CNN-LSTM, dengan CNN berperan sebagai ekstraktor fitur visual dari setiap *frame*, sedangkan LSTM digunakan untuk mempelajari keterkaitan fitur antarframe sebagai rangkaian informasi visual. Untuk mengurangi redundansi informasi antarframe dan risiko kemiripan berlebihan antara data latih dan data uji, jumlah frame dibatasi maksimum 15 *frame* per folder video serta pembagian data dilakukan dengan memperhatikan kelompok sumber video. Dataset terdiri dari bangunan Melayu Riau, bangunan Melayu Kalimantan, dan bangunan non-melayu. Tahapan penelitian meliputi ekstraksi frame, resize citra menjadi 224×224 piksel, normalisasi, augmentasi data, pelabelan kelas, pembagian data latih dan data uji, pemodelan, evaluasi, validasi menggunakan GroupKFold, serta implementasi sistem berbasis *web*. Hasil pengujian menunjukkan bahwa skenario terbaik diperoleh pada *split* data 80:20, *epoch* maksimum 80, dan *batch size* 16 dengan *accuracy* sebesar 0,9916 dan *test loss* sebesar 0,0852. Validasi GroupKFold menghasilkan rata-rata *accuracy* sebesar 99,1% dengan standar deviasi 0,5%. Hasil ini menunjukkan bahwa model mampu mengenali pola visual arsitektur, seperti atap, jendela, pintu, ornamen, dan tampilan keseluruhan bangunan, namun performa tersebut tetap perlu dipahami dalam ruang lingkup dataset dan skenario evaluasi yang digunakan.

Kata Kunci: Analisis Video; Bangunan Melayu; CNN-LSTM; Deep Learning; Klasifikasi Bangunan

Abstract−Malay buildings have distinctive architectural characteristics and require technology-based identification systems to support cultural documentation and preservation. This study aims to develop an identification system for Malay and non-Malay buildings using video data extracted into frame-by-frame images. The use of video data in this study is not intended to analyze the physical movement of buildings, but to utilize visual variations caused by changes in camera angle, recording distance, lighting, object composition, and visible building elements. The proposed method is CNN-LSTM, where CNN extracts visual features from each frame, while LSTM learns inter-frame feature relationships as a sequence of visual information. To reduce redundant information between adjacent frames and minimize the risk of excessive similarity between training and testing data, the number of frames was limited to a maximum of 15 frames per video folder, and data splitting was performed by considering video source groups. The dataset consists of Riau Malay buildings, Kalimantan Malay buildings, and non-Malay buildings. The research stages include frame extraction, image resizing to 224×224 pixels, normalization, data augmentation, class labeling, train-test splitting, modeling, evaluation, GroupKFold validation, and web-based system implementation. The best testing scenario was obtained using an 80:20 data split, 80 maximum epochs, and a batch size of 16, achieving an accuracy of 0.9916 and a test loss of 0.0852. GroupKFold validation produced an average accuracy of 99.1% with a standard deviation of 0.5%. These results indicate that the model can recognize architectural visual patterns, such as roofs, windows, doors, ornaments, and overall building appearance, while the performance should still be interpreted within the scope of the dataset and evaluation scenario used in this study.

Keywords: Video Analysis; Malay Buildings; CNN-LSTM; Deep Learning; Building Classification

1. PENDAHULUAN

Perkembangan teknologi digital, khususnya dalam bidang *computer vision* dan *artificial intelligence*, telah memberikan dampak signifikan dalam berbagai aspek kehidupan, termasuk dalam pelestarian budaya [1], [2]. Pemanfaatan teknologi ini memungkinkan proses dokumentasi, analisis, dan identifikasi objek budaya dilakukan secara lebih sistematis dan efisien. Salah satu bentuk warisan budaya yang memiliki nilai historis dan identitas lokal yang kuat adalah bangunan tradisional Melayu yang banyak ditemukan di wilayah Riau dan Kalimantan. Bangunan ini memiliki ciri khas berupa struktur panggung, bentuk atap yang khas, serta ornamen ukiran yang sarat nilai filosofis. Namun demikian, seiring dengan pesatnya modernisasi dan urbanisasi, keberadaan bangunan Melayu semakin berkurang dan menghadapi ancaman kepunahan, baik dari segi fisik maupun dokumentasi digital [3], [4]. Oleh karena itu, diperlukan upaya pelestarian berbasis teknologi yang mampu menjaga keberlangsungan warisan budaya tersebut.

Keterbatasan dalam proses dokumentasi dan identifikasi bangunan Melayu menjadi permasalahan utama dalam upaya pelestarian budaya. Saat ini, proses identifikasi masih dilakukan secara manual oleh ahli arsitektur melalui observasi visual langsung, yang tidak hanya membutuhkan waktu lama tetapi juga bergantung pada subjektivitas individu sehingga berpotensi menghasilkan inkonsistensi dalam penilaian [5]. Selain itu, pendekatan manual sulit diterapkan dalam skala besar, terutama ketika jumlah objek yang harus dianalisis semakin banyak dan tersebar di berbagai wilayah geografis.

Building of Informatics, Technology and Science (BITS)
Volume 8, No 1, June 2026 Page: 318−329
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v8i1.9917

Di sisi lain, kurangnya pemanfaatan teknologi dalam mengolah data visual seperti citra dan video menyebabkan potensi digitalisasi warisan budaya belum dimaksimalkan secara optimal [3], [6]. Ketersediaan dataset visual yang terbatas, tidak terstandar, serta minimnya anotasi data juga menjadi hambatan dalam pengembangan sistem berbasis kecerdasan buatan untuk mengidentifikasi bangunan tradisional [7]. Oleh karena itu, diperlukan suatu pendekatan berbasis teknologi yang mampu mengotomatisasi proses identifikasi bangunan Melayu secara lebih cepat, objektif, dan efisien dengan memanfaatkan kemajuan *computer vision* dan *deep learning* [3], [8].

Dalam upaya mengatasi permasalahan tersebut, *machine learning* dan *deep learning* telah banyak digunakan dalam pengolahan data visual karena kemampuannya dalam mempelajari pola dari data secara otomatis [2], [8]. Metode *Convolutional Neural Network* (CNN) merupakan salah satu algoritma yang paling populer dalam klasifikasi citra karena kemampuannya dalam mengekstraksi fitur visual seperti bentuk, tekstur, dan pola secara otomatis dengan tingkat akurasi yang tinggi [2], [9]. CNN telah berhasil diterapkan dalam berbagai penelitian, termasuk klasifikasi objek, deteksi wajah, serta identifikasi bangunan bersejarah [10].

Pada data video bangunan, analisis visual tidak hanya berkaitan dengan satu citra tunggal, tetapi juga dengan rangkaian *frame* yang merepresentasikan objek dari beberapa tampilan. Meskipun bangunan merupakan objek diam, perubahan posisi kamera selama proses perekaman dapat menghasilkan variasi tampilan visual, seperti perbedaan sudut pengambilan gambar, jarak kamera, pencahayaan, komposisi objek, serta bagian bangunan yang terekam. Variasi tersebut penting karena satu citra tunggal belum tentu memperlihatkan seluruh karakteristik bangunan secara utuh, terutama pada elemen seperti atap, jendela, pintu, ornamen, dan tampilan keseluruhan bangunan. Oleh karena itu, penggunaan *frame* dari video dapat memberikan representasi visual yang lebih lengkap dalam proses identifikasi bangunan Melayu. Hal ini sejalan dengan penelitian yang menunjukkan bahwa penggunaan beberapa tampilan atau sudut pandang objek dapat membantu memperoleh representasi visual yang lebih lengkap dibandingkan hanya menggunakan satu tampilan [11], [12].

CNN memiliki kemampuan yang baik dalam mengekstraksi fitur spasial dari citra, tetapi proses tersebut umumnya dilakukan terhadap setiap frame sebagai data visual yang berdiri sendiri. Pada data video bangunan, setiap frame dapat memuat informasi visual yang saling melengkapi. Misalnya, satu *frame* dapat lebih menonjolkan bentuk atap, sedangkan *frame* lain memperlihatkan jendela, pintu, ornamen, atau tampilan keseluruhan bangunan. Apabila setiap *frame* hanya diproses secara terpisah dan hasilnya digabungkan pada tahap akhir, keterkaitan visual antarframe belum dimanfaatkan secara langsung oleh model. Oleh karena itu, diperlukan pendekatan yang mampu mempelajari fitur visual dari setiap *frame* sekaligus mempertimbangkan hubungan berurutan antarframe sebagai satu rangkaian informasi visual [13], [14].

Untuk mengolah data visual berurutan tersebut, digunakan metode *Long Short-Term Memory* (LSTM), yang merupakan pengembangan dari *Recurrent Neural Network* (RNN) dan dirancang khusus untuk menangani data berurutan (*sequential* data). LSTM memiliki mekanisme gating yang memungkinkan model untuk mengontrol aliran informasi serta mempertahankan informasi penting dalam jangka panjang [15], [16]. Pada konteks data video bangunan, LSTM tidak digunakan untuk menganalisis pergerakan objek bangunan, melainkan untuk mempelajari keterkaitan fitur visual antarframe yang terbentuk akibat perubahan sudut pandang kamera selama proses perekaman. Fitur spasial yang diekstraksi CNN dari setiap *frame* kemudian diperlakukan sebagai urutan fitur, sehingga model dapat mempertimbangkan hubungan visual antarframe selain informasi pada citra tunggal [15], [17].

Integrasi antara CNN dan LSTM sebagai metode *hybrid* CNN-LSTM memungkinkan sistem untuk tidak hanya mengekstraksi fitur spasial dari setiap *frame* menggunakan CNN, tetapi juga menganalisis keterkaitan antar *frame* menggunakan LSTM, sehingga menghasilkan representasi data yang lebih komprehensif [13]. Pendekatan ini terbukti meningkatkan performa dalam berbagai tugas klasifikasi berbasis video dibandingkan penggunaan model tunggal [17], [18]. Pada data video bangunan, rangkaian frame tidak dipahami sebagai urutan gerak objek, tetapi sebagai variasi tampilan visual yang muncul karena perubahan sudut pandang, jarak pengambilan gambar, pencahayaan, dan komposisi objek akibat pergerakan kamera. Oleh karena itu, penggunaan metode hybrid CNN-LSTM menjadi relevan karena CNN mengekstraksi fitur spasial dari setiap *frame*, sedangkan LSTM mempelajari keterkaitan fitur antarframe yang merepresentasikan variasi tampilan bangunan [13], [19].

Berbagai penelitian telah menerapkan metode CNN-LSTM dalam berbagai domain. Penelitian oleh Tipper *et al*. menunjukkan bahwa CNN-LSTM mampu digunakan untuk mendeteksi *deepfake* video dengan tingkat akurasi tinggi [20]. Penelitian oleh Khan dan Jung juga menerapkan CNN-LSTM untuk human *activity recognition* dan memperoleh peningkatan performa dibandingkan metode konvensional [18]. Selain itu, Pandey *et al*. dalam studinya menyatakan bahwa model *hybrid* CNN-LSTM memiliki keunggulan dalam menangani data sekuensial dibandingkan model tunggal [13]. Selanjutnya, Mao *et al*. menegaskan bahwa pendekatan *spatio-temporal* modeling menggunakan CNN-LSTM merupakan metode yang efektif dalam klasifikasi video modern [14].

Dalam bidang arsitektur, penelitian oleh Mishra *et al*. menunjukkan bahwa CNN mampu mengklasifikasikan gaya bangunan dengan baik berdasarkan pola visual yang khas [10]. Penelitian lain oleh Tasir *et al*. menunjukkan bahwa CNN dapat digunakan untuk mengidentifikasi bangunan bersejarah dengan tingkat akurasi yang tinggi [9]. Sementara itu, Alzahrani *et al.* menjelaskan bahwa penggunaan beberapa tampilan atau sudut pandang objek dapat membantu memperoleh representasi visual yang lebih lengkap dibandingkan hanya menggunakan satu tampilan [11]. Temuan tersebut mendukung penggunaan data video pada penelitian ini, karena rangkaian *frame* video bangunan dapat merepresentasikan variasi sudut pandang kamera yang tidak selalu diperoleh dari satu citra tunggal.

Building of Informatics, Technology and Science (BITS)
Volume 8, No 1, June 2026 Page: 318−329
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v8i1.9917

Berdasarkan penelitian-penelitian tersebut, dapat disimpulkan bahwa CNN dan LSTM telah banyak digunakan dalam klasifikasi data visual dan sekuensial. Namun, sebagian besar penelitian masih berfokus pada domain seperti aktivitas manusia, keamanan video, dan klasifikasi objek umum. Dalam bidang arsitektur, penelitian terdahulu menunjukkan bahwa CNN mampu digunakan untuk klasifikasi citra bangunan atau identifikasi bangunan bersejarah berbasis citra statis [9], [10]. Akan tetapi, pemanfaatan data video untuk identifikasi bangunan tradisional masih belum banyak dikaji, khususnya dalam memanfaatkan variasi tampilan visual antarframe yang terbentuk akibat perubahan perspektif kamera selama proses perekaman. Dengan demikian, kesenjangan penelitian ini terletak pada masih terbatasnya pemanfaatan rangkaian *frame* video sebagai sumber informasi visual berurutan untuk identifikasi bangunan Melayu. Oleh karena itu, metode hybrid CNN-LSTM digunakan agar CNN dapat mengekstraksi fitur visual dari setiap *frame*, sedangkan LSTM mempelajari keterkaitan fitur antarframe sebagai rangkaian informasi visual yang merepresentasikan variasi tampilan bangunan Melayu selama proses perekaman [12], [14], [15].

Penelitian ini bertujuan untuk mengembangkan sistem identifikasi bangunan Melayu menggunakan metode CNN-LSTM dengan memanfaatkan data video sebagai *input*. Sistem ini diharapkan mampu meningkatkan akurasi identifikasi dengan memanfaatkan informasi spasial dari setiap *frame* serta keterkaitan visual antarframe secara bersamaan. Selain itu, penelitian ini juga bertujuan untuk mengevaluasi performa model menggunakan metrik *accuracy, precision, recall, F1-score*, dan *test loss.*

Dalam proses evaluasi, digunakan beberapa skenario pengujian berdasarkan variasi pembagian data, yaitu rasio 70:30 dan 80:20, serta variasi jumlah epoch maksimum dan *batch size* untuk mengetahui konfigurasi model yang menghasilkan performa paling optimal dalam melakukan klasifikasi bangunan Melayu dan non-Melayu. Selain itu, dilakukan validasi tambahan menggunakan K-Fold *Cross-Validation* dengan metode GroupKFold untuk mengetahui kestabilan performa model pada beberapa pembagian data serta mengurangi risiko bias akibat pemilihan data tertentu. Dengan demikian, penelitian ini diharapkan dapat memberikan kontribusi dalam pengembangan teknologi *computer vision* serta mendukung pelestarian budaya melalui digitalisasi arsitektur tradisional.

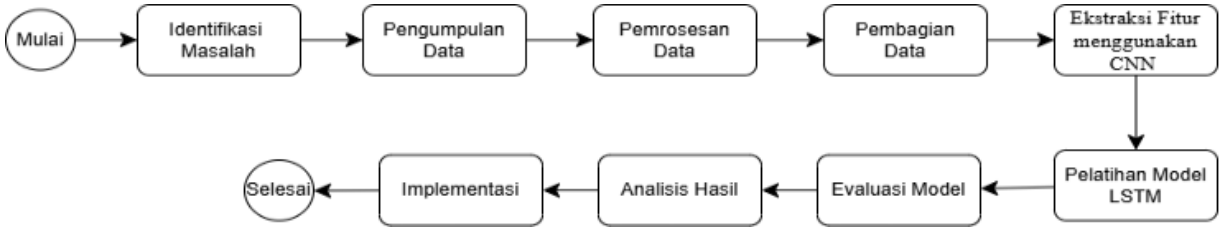
2. METODOLOGI PENELITIAN

2.1 Objek Penelitian

Objek penelitian ini adalah bangunan yang diklasifikasikan ke dalam tiga kategori utama, yaitu Bangunan Melayu Riau, Bangunan Melayu Kalimantan, dan bangunan non-melayu. Bangunan Melayu Riau dan Bangunan Melayu Kalimantan digunakan untuk merepresentasikan karakteristik arsitektur Melayu berdasarkan wilayah asalnya, sedangkan bangunan non-melayu digunakan sebagai kelas pembanding agar model mampu membedakan bangunan yang memiliki dan tidak memiliki karakteristik arsitektur Melayu. Pada kategori Bangunan Melayu Riau dan Bangunan Melayu Kalimantan, data disusun berdasarkan elemen visual bangunan, seperti atap, jendela, pintu, ornamen, dan tampilan keseluruhan bangunan. Elemen tersebut digunakan karena memiliki karakteristik visual yang dapat menjadi pembeda arsitektur, baik dari segi bentuk, pola, tekstur, maupun struktur bangunan. Sementara itu, kelas non-melayu digunakan untuk memperkuat kemampuan model dalam mengenali objek yang tidak termasuk dalam kategori bangunan Melayu.

2.2 Tahapan Penelitian

Penelitian ini menggunakan metode eksperimen dengan pendekatan komputasional berbasis *deep learning*. Tahapan penelitian dilakukan secara sistematis mulai dari identifikasi masalah, pengumpulan data, *preprocessing* data, pembagian data, ekstraksi fitur menggunakan CNN, pelatihan model LSTM, evaluasi model, analisis hasil, hingga implementasi sistem berbasis *web*. Pengujian dilakukan melalui beberapa skenario berdasarkan variasi *split* data, *epoch* maksimum, dan *batch size* untuk mengetahui konfigurasi model terbaik. Validasi tambahan menggunakan GroupKFold juga dilakukan untuk mengetahui kestabilan performa model pada beberapa pembagian data. Alur tahapan penelitian ditunjukkan pada Gambar 1.



Gambar 1. Tahapan Penelitian

a. Identifikasi Masalah
Tahap ini dilakukan untuk merumuskan permasalahan penelitian, yaitu belum optimalnya pemanfaatan pendekatan *deep learning* dalam mengidentifikasi bangunan Melayu secara otomatis berdasarkan data video. Penelitian ini difokuskan pada klasifikasi bangunan Melayu Riau, bangunan Melayu Kalimantan, dan bangunan

Building of Informatics, Technology and Science (BITS)
Volume 8, No 1, June 2026 Page: 318−329
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v8i1.9917

non-melayu. Identifikasi dilakukan dengan memanfaatkan karakteristik visual bangunan yang tampak pada *frame* video, baik pada bagian atap, jendela, pintu, ornamen, maupun tampilan keseluruhan bangunan.

b. Pengumpulan Data
Data yang digunakan berupa video bangunan Melayu Riau, bangunan Melayu Kalimantan, dan bangunan non-melayu. Data disusun berdasarkan kelas masing-masing agar proses pelabelan dapat dilakukan secara konsisten. Data video dikumpulkan dari beberapa objek bangunan dan kondisi perekaman yang berbeda. Perbedaan tersebut mencakup sudut pengambilan gambar, jarak kamera, pencahayaan, komposisi objek, bagian bangunan yang terekam, dan latar visual. Variasi ini penting agar model tidak hanya mempelajari ciri visual yang melekat pada satu latar atau satu kondisi perekaman, tetapi juga mampu mengenali karakteristik arsitektur bangunan Melayu secara lebih umum.

c. *Preprocessing* Data
Preprocessing bertujuan untuk meningkatkan kualitas data sebelum digunakan dalam pelatihan model. Langkah-langkah *preprocessing* meliputi:
1. Ekstraksi *Frame*
Data video diubah menjadi kumpulan *frame* yang merepresentasikan tampilan visual bangunan. Jumlah *frame* dari setiap folder video dibatasi maksimum 15 *frame* sebagai strategi *sampling* untuk mengurangi redundansi visual dan menjaga keseimbangan kontribusi data antarvideo. Pembatasan ini dilakukan karena *frame* dari video yang sama cenderung memiliki kemiripan tinggi dari segi objek, sudut kamera, latar belakang, dan pencahayaan [21]. Batas 15 *frame* digunakan agar video berdurasi panjang tidak mendominasi proses pelatihan, sekaligus menjaga efisiensi komputasi dan keterwakilan visual dari setiap video.
2. *Resize* dan Normalisasi
Seluruh citra diubah ukurannya menjadi 224×224 piksel agar sesuai dengan kebutuhan *input* model CNN. Selanjutnya, dilakukan normalisasi nilai piksel ke dalam rentang [-1, 1] untuk menjaga konsistensi skala data dan mempercepat proses pelatihan model.
3. Augmentasi Data
Augmentasi dilakukan pada data latih melalui rotasi, *flipping* horizontal, dan penyesuaian posisi citra untuk meningkatkan variasi data serta mengurangi risiko *overfitting*. Augmentasi hanya diterapkan pada data latih, sedangkan data uji tidak diberi augmentasi agar evaluasi tetap objektif.
4. *Labeling*
Setiap citra diberikan label berdasarkan kategori yang telah ditentukan, yang dilakukan secara otomatis sesuai dengan struktur folder dataset.

d. Pembagian Data
Pembagian data dilakukan untuk memisahkan dataset menjadi data latih dan data uji menggunakan dua rasio, yaitu 70:30 dan 80:20. Variasi rasio ini digunakan untuk mengetahui pengaruh proporsi data latih dan data uji terhadap performa model. Selain itu, pengujian juga dilakukan dengan beberapa kombinasi *epoch* maksimum dan *batch size* untuk menentukan konfigurasi pelatihan terbaik. Pembagian data dilakukan menggunakan GroupShuffleSplit dengan memperhatikan sumber video sebagai kelompok data, sehingga *frame* yang berasal dari video atau lokasi yang sama tidak masuk secara bersamaan ke data latih dan data uji. Hal ini dilakukan karena *frame* dari satu video cenderung memiliki kemiripan visual tinggi, baik dari segi latar belakang, sudut kamera, maupun pencahayaan, sehingga risiko data *leakage* dapat dikurangi. Untuk memperkuat evaluasi, penelitian ini juga menggunakan GroupKFold *5-fold cross-validation* guna mengetahui kestabilan performa model pada beberapa pembagian data.

e. Ekstraksi Fitur menggunakan CNN
Ekstraksi fitur dilakukan untuk mengubah citra menjadi representasi numerik yang dapat diproses oleh model. Pada penelitian ini, MobileNetV2 digunakan sebagai backbone CNN karena memiliki arsitektur ringan dan efisien, sehingga sesuai untuk memproses citra hasil ekstraksi *frame* serta mendukung implementasi sistem berbasis *web*. Pemilihan MobileNetV2 tidak dimaksudkan untuk menyatakan bahwa arsitektur ini lebih unggul dibandingkan arsitektur yang lebih dalam, tetapi didasarkan pada keseimbangan antara kemampuan ekstraksi fitur, efisiensi komputasi, dan kebutuhan implementasi sistem. Meskipun termasuk arsitektur ringan, model digunakan melalui pendekatan *transfer learning* yang umum dimanfaatkan pada klasifikasi citra warisan budaya dengan dataset terbatas [22]. Pada input citra 224×224 piksel, *backbone* CNN menghasilkan *feature map* berukuran 7×7×1280. Representasi fitur tersebut diringkas melalui *row-wise average pooling* menjadi 7×1280, kemudian direduksi menjadi 7×128 sebelum diproses oleh LSTM. Dengan demikian, penggunaan 7 *timestep* didasarkan pada struktur keluaran fitur dari CNN.

f. Pelatihan Model *Long Short Term Memory* (LSTM)
Pelatihan model dilakukan menggunakan *Long Short Term Memory* (LSTM) untuk mempelajari pola sekuensial dari vektor fitur yang telah diekstraksi. Data disusun dalam bentuk *sequence* untuk merepresentasikan keterkaitan visual antarframe. LSTM dipilih karena kemampuannya dalam menangkap dependensi jangka panjang pada data sekuensial [15]. Selain itu, LSTM juga mampu mengatasi permasalahan *vanishing gradient* yang sering terjadi pada model *Recurrent Neural Network* (RNN) konvensional [23]. Dalam penelitian ini, model dilatih menggunakan data latih dengan menerapkan regularisasi seperti *dropout* dan L2 untuk mengurangi *overfitting* dan meningkatkan kemampuan generalisasi.

g. Evaluasi Model

Building of Informatics, Technology and Science (BITS)
Volume 8, No 1, June 2026 Page: 318−329
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v8i1.9917

Setelah proses pelatihan selesai, dilakukan evaluasi untuk mengukur kinerja model dalam memprediksi data baru. Evaluasi dilakukan menggunakan metrik *accuracy, precision, recall, F1-score,* dan *test loss*. *Accuracy* digunakan untuk mengukur ketepatan prediksi secara keseluruhan; precision menunjukkan ketepatan model dalam memprediksi suatu kelas; recall menunjukkan kemampuan model dalam mengenali seluruh data pada kelas tertentu; sedangkan *F1-score* merupakan kombinasi harmonis antara *precision* dan *recall*. Selain itu, digunakan *confusion matrix* untuk memberikan gambaran lebih rinci mengenai performa model melalui empat komponen utama, yaitu *True Positive, False Positive, True Negative,* dan *False Negative* [24]. Evaluasi juga diperkuat dengan *K-Fold Cross-Validation* menggunakan GroupKFold sebanyak 5 fold untuk mengetahui kestabilan performa model dan mengurangi risiko bias akibat pembagian data tertentu. Hasil evaluasi ini digunakan untuk menilai kemampuan model dalam mengenali pola data serta menghasilkan prediksi yang akurat sebelum dilakukan analisis lebih lanjut.

h. Analisis Hasil
Tahap ini dilakukan untuk menganalisis hasil evaluasi model CNN-LSTM berdasarkan metrik kinerja yang diperoleh. Analisis difokuskan pada perbandingan performa setiap skenario pengujian, pemilihan konfigurasi terbaik, kestabilan hasil validasi menggunakan GroupKFold, serta pola kesalahan prediksi yang terlihat pada *confusion matrix*. Berdasarkan analisis tersebut, dapat diketahui efektivitas model dalam menangkap karakteristik spasial dan keterkaitan visual antarframe dari data video serta faktor-faktor yang memengaruhi performa klasifikasi bangunan tradisional Melayu.

i. Implementasi
Tahap implementasi dilakukan dengan membangun sistem identifikasi bangunan tradisional Melayu berbasis *web* menggunakan *framework* Flask yang terintegrasi dengan model CNN-LSTM hasil pelatihan. Sistem dirancang untuk menerima input berupa gambar maupun video, kemudian melakukan proses *preprocessing*, ekstraksi fitur visual, serta klasifikasi berdasarkan fitur spasial dan keterkaitan visual antarframe menggunakan LSTM untuk menentukan kategori bangunan secara otomatis. Hasil identifikasi pada sistem ditampilkan sebagai kategori akhir berupa Bangunan Melayu Riau, Bangunan Melayu Kalimantan, atau non-melayu. Penyajian hasil dalam bentuk kategori utama ini bertujuan agar informasi klasifikasi lebih mudah dipahami oleh pengguna melalui antarmuka *web*.

2.3 Metode *Hybrid* CNN-LSTM

2.3.1 *Convolutional Neural Network* (CNN)

Convolutional Neural Network (CNN) merupakan salah satu metode *deep learning* yang banyak digunakan dalam bidang *computer vision* karena kemampuannya dalam mengekstraksi fitur visual secara otomatis dari data citra. CNN bekerja melalui beberapa lapisan utama, yaitu *convolution layer, activation layer*, dan *pooling layer,* yang secara bertahap mempelajari pola visual mulai dari tepi, tekstur, bentuk, hingga karakteristik objek yang lebih kompleks [25]. Dibandingkan metode ekstraksi fitur konvensional, CNN mampu menghasilkan representasi fitur yang lebih *robust* sehingga banyak diterapkan pada berbagai penelitian klasifikasi citra [26]. Secara matematis, proses konvolusi pada CNN sebagai berikut:

$$y_{i,j} = \sum_m \sum_n x_{i+m,j+n} w_{m,n} + b \quad (1)$$

Dengan x sebagai *input* citra, w sebagai kernel konvolusi, b sebagai bias, dan y sebagai keluaran berupa *feature map*. Operasi konvolusi tersebut memungkinkan model mengenali pola spasial lokal pada citra secara otomatis sehingga informasi visual yang relevan dapat direpresentasikan dalam bentuk fitur berdimensi tinggi [27]. Selain itu, proses *pooling* digunakan untuk mereduksi dimensi fitur sekaligus mempertahankan informasi penting, sehingga meningkatkan efisiensi komputasi model [28]. Pada penelitian ini, CNN berperan dalam mengekstraksi karakteristik visual bangunan tradisional Melayu, seperti bentuk atap, struktur bangunan, dan ornamen khas arsitektur, sehingga menghasilkan representasi fitur yang selanjutnya digunakan pada tahap pemodelan sekuensial.

2.3.2 *Long Short-Term Memory* (LSTM)

Long Short-Term Memory (LSTM) merupakan pengembangan dari *Recurrent Neural Network* (RNN) yang dirancang untuk memproses data berurutan (*sequential* data) dengan kemampuan mempertahankan informasi dalam jangka panjang. LSTM memiliki mekanisme *memory cell* yang dikendalikan oleh beberapa komponen utama, yaitu *forget gate, input gate,* dan *output gate*, sehingga mampu menyimpan informasi penting dan mengurangi permasalahan *vanishing gradient* yang umum terjadi pada RNN konvensional [25]. Karakteristik tersebut menjadikan LSTM efektif digunakan dalam pemodelan data temporal maupun data visual berurutan [26]. Fungsi *forget gate* pada LSTM dirumuskan sebagai:

$$f_t = \sigma\left(W_f . [h_{t-1}, x_t] + b_f\right) \quad (2)$$

Sedangkan pembaruan *cell state* dinyatakan dengan persamaan:

$$C_t = f_t * C_{t-1} + i_t * C'_t \quad (3)$$

Building of Informatics, Technology and Science (BITS)
Volume 8, No 1, June 2026 Page: 318−329
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v8i1.9917

Di mana f_t merupakan *forget gate*, i_t merupakan *input gate,* C_t adalah *cell state,* dan C_t' merupakan kandidat memori baru. Mekanisme ini memungkinkan LSTM mempertahankan informasi yang relevan dari urutan data sebelumnya untuk digunakan pada proses prediksi berikutnya [27]. Dalam penelitian ini, LSTM digunakan untuk mempelajari hubungan antar representasi fitur hasil ekstraksi CNN, sehingga model mampu memahami keterkaitan pola visual secara berurutan sebelum dilakukan proses klasifikasi akhir. Secara integratif, metode *Hybrid* CNN-LSTM memanfaatkan CNN sebagai *feature extractor* untuk menghasilkan representasi fitur spasial, kemudian LSTM memproses urutan fitur tersebut guna mempelajari keterkaitan visual antarframe sebelum dilakukan klasifikasi akhir menggunakan fungsi *Softmax*, yang dirumuskan sebagai:

$$P(y_i) = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}} \tag{4}$$

Dengan $P(y_i)$ merupakan probabilitas kelas ke-i, z_i adalah nilai aktivasi kelas, dan K adalah jumlah kelas. Kelas dengan probabilitas tertinggi dipilih sebagai hasil prediksi akhir. Kombinasi ini memungkinkan model memanfaatkan informasi spasial dan keterkaitan visual antarframe secara bersamaan sehingga menghasilkan performa klasifikasi yang lebih optimal [26], [28].

3. HASIL DAN PEMBAHASAN

3.1 Persiapan Data

Dataset yang digunakan dalam penelitian ini berasal dari data video bangunan yang telah diekstraksi menjadi citra *frame-by-frame*. Data video tidak digunakan secara langsung dalam proses pelatihan model, melainkan terlebih dahulu diubah menjadi kumpulan *frame* agar dapat diproses sebagai data citra oleh model CNN–LSTM. Proses ekstraksi *frame* dilakukan untuk memperoleh representasi visual bangunan dari beberapa sudut pandang, sehingga model dapat mempelajari karakteristik visual bangunan secara lebih detail. Pendekatan berbasis *frame* video relevan digunakan dalam pemrosesan visual karena CNN dapat mengekstraksi fitur spasial dari setiap *frame*, sedangkan LSTM dapat mempelajari keterkaitan visual antarframe dari fitur yang dihasilkan [29].

Dataset pada penelitian ini terdiri dari dua kelompok utama, yaitu bangunan Melayu dan bangunan non-Melayu. Data bangunan Melayu merepresentasikan karakteristik arsitektur dari wilayah Riau dan Kalimantan, sedangkan data non-Melayu digunakan untuk membantu model membedakan bangunan Melayu dengan bangunan lain yang tidak memiliki karakteristik arsitektur Melayu. Penambahan kelas non-melayu penting dilakukan karena pada kondisi implementasi nyata, pengguna dapat mengunggah citra bangunan yang tidak termasuk dalam kategori bangunan Melayu. Dengan demikian, sistem tidak hanya berfungsi untuk mengenali bangunan Melayu, tetapi juga mampu memberikan pembeda terhadap objek bangunan yang tidak memiliki karakteristik arsitektur Melayu.

Dataset disusun ke dalam beberapa kelas, yaitu kalimantan_atap, kalimantan_jendela, Kalimantan keseluruhan, kalimantan_ornamen, kalimantan_pintu, riau_atap, riau_jendela, riau_keseluruhan, riau_ornamen, dan non_melayu. Pembagian kelas tersebut bertujuan agar model mampu mengenali elemen-elemen visual penting pada bangunan, seperti atap, jendela, pintu, ornamen, dan tampilan keseluruhan bangunan. Elemen-elemen tersebut dipilih karena memiliki ciri visual yang dapat menjadi pembeda antara bangunan Melayu Riau, bangunan Melayu Kalimantan, dan bangunan non-Melayu. Contoh sampel citra dari masing-masing kelompok data yang digunakan dalam penelitian ini ditunjukkan pada Tabel 1.

Tabel 1. Dataset Bangunan Melayu dan Non-Melayu

Kelompok Data	Sampel Citra		
Melayu Kalimantan	 ezgif-frame-001.jpg	 ezgif-frame-002.jpg	 ezgif-frame-003.jpg
Melayu Riau	 ezgif-frame-001.jpg	 ezgif-frame-002.jpg	 ezgif-frame-003.jpg
Non-Melayu	 ezgif-frame-001.jpg	 ezgif-frame-002.jpg	 ezgif-frame-003.jpg

Building of Informatics, Technology and Science (BITS)
Volume 8, No 1, June 2026 Page: 318−329
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v8i1.9917

Tahapan prapemrosesan data diawali dengan ekstraksi *frame* dari video. Setiap video diproses menjadi beberapa *frame* yang merepresentasikan tampilan visual bangunan. Karena *frame* yang berasal dari video yang sama memiliki tingkat kemiripan visual yang tinggi, pembagian data perlu dilakukan dengan memperhatikan sumber video. Hal ini dilakukan untuk mengurangi risiko data *leakage*. Data *leakage* dapat terjadi apabila *frame* dari video yang sama masuk secara bersamaan ke dalam data latih dan data uji. Apabila hal tersebut terjadi, hasil evaluasi model dapat menjadi terlalu tinggi karena data uji memiliki kemiripan yang sangat dekat dengan data latih. Oleh karena itu, pembagian data dilakukan dengan mempertimbangkan kelompok sumber video agar evaluasi model menjadi lebih objektif.

Setelah *frame* diperoleh, seluruh citra diubah ukurannya menjadi 224×224 piksel agar sesuai dengan ukuran *input* model. Selanjutnya, dilakukan normalisasi nilai piksel untuk menyamakan skala data sebelum digunakan dalam proses pelatihan model. Normalisasi diperlukan agar nilai piksel berada pada rentang yang lebih stabil sehingga proses pembelajaran model dapat berjalan lebih baik. Selain itu, augmentasi data dilakukan untuk meningkatkan variasi data latih dan mengurangi risiko *overfitting*. Teknik augmentasi yang digunakan meliputi rotasi, *flipping horizontal*, dan penyesuaian posisi citra.

Setelah seluruh data siap, dataset dibagi ke dalam dua rasio, yaitu 70:30 dan 80:20, untuk keperluan pelatihan dan evaluasi model. Pembagian data dilakukan untuk mengetahui pengaruh proporsi data latih dan data uji terhadap performa model *hybrid* CNN-LSTM. Pada setiap skenario rasio, model dilatih menggunakan data latih dan dievaluasi menggunakan data uji, sehingga hasil evaluasi dapat menunjukkan kemampuan model dalam mengenali data baru. Selain itu, dilakukan variasi epoch maksimum dan *batch size* untuk mengetahui pengaruh konfigurasi pelatihan terhadap performa model.

Seluruh tahapan prapemrosesan data menghasilkan dataset yang bersih, lengkap, dan terstruktur dengan baik. Dataset ini terdiri dari kelas target yang merepresentasikan bangunan Melayu Kalimantan, bangunan Melayu Riau, dan bangunan non-Melayu. Setelah melalui proses ekstraksi *frame*, *resize*, normalisasi, augmentasi, *labeling*, dan pembagian data, dataset berada dalam format citra yang siap digunakan untuk melatih model CNN-LSTM. Untuk memastikan kestabilan performa model, dilakukan pula validasi tambahan menggunakan GroupKFold sebanyak 5 fold dengan mempertimbangkan kelompok sumber video.

3.2 Pemodelan dan Evaluasi

Setelah dataset siap digunakan, tahap selanjutnya adalah proses pemodelan menggunakan arsitektur CNN-LSTM. Model ini dipilih karena mampu menggabungkan proses ekstraksi fitur visual dari citra dengan pembelajaran keterkaitan visual antarframe dari fitur hasil ekstraksi. CNN digunakan sebagai ekstraktor fitur dari setiap *frame* bangunan, sedangkan LSTM digunakan untuk mempelajari hubungan antarfitur yang telah disusun dalam bentuk urutan. Dengan pendekatan ini, model diharapkan mampu mengenali pola visual bangunan Melayu dan non-Melayu secara lebih optimal.

Pengujian model dilakukan menggunakan beberapa skenario berdasarkan variasi *split* data, *epoch* maksimum, dan *batch size*. Variasi *split* data menggunakan dua rasio, yaitu 70:30 dan 80:20, untuk mengetahui pengaruh proporsi data latih dan data uji terhadap performa model. Sementara itu, variasi *epoch* maksimum dan *batch size* digunakan untuk mengetahui pengaruh konfigurasi pelatihan terhadap hasil evaluasi. *Learning rate* awal yang digunakan pada setiap skenario adalah 1e-4, sedangkan *learning rate* pada tahap *fine-tuning* adalah 1e-5. Penggunaan *learning rate* yang lebih kecil pada tahap *fine-tuning* bertujuan agar bobot model dapat menyesuaikan diri terhadap dataset penelitian secara lebih stabil tanpa menyebabkan perubahan bobot yang terlalu besar. Hasil pengujian model CNN-LSTM berdasarkan variasi *split* data, *epoch* maksimum, dan *batch size* ditunjukkan pada Tabel 2.

Tabel 2. Perbandingan Hasil Pengujian Berdasarkan *Split* Data, *Epoch*, dan *Batch Size*

Model	Split Data	*Epoch Maks*	*Batch Size*	*Learning Rate*	*Fine-Tuning Learning Rate*	*Accuracy*	*Test Loss*
CNN-LSTM	70:30	50	16	1e-4	1e-5	0,9888	0,0869
CNN-LSTM	70:30	50	32	1e-4	1e-5	0,9888	0,1002
CNN-LSTM	70:30	80	16	1e-4	1e-5	0,9898	0,1130
CNN-LSTM	80:20	50	16	1e-4	1e-5	0,9874	0,0881
CNN-LSTM	80:20	50	32	1e-4	1e-5	0,9916	0,0924
CNN-LSTM	80:20	80	16	1e-4	1e-5	0,9916	0,0852

Berdasarkan Tabel 2, nilai *accuracy* tertinggi sebesar 0,9916 diperoleh pada dua skenario, yaitu *split* data 80:20 dengan *epoch* maksimum 50 dan *batch size* 32, serta *split* data 80:20 dengan *epoch* maksimum 80 dan *batch size* 16. Meskipun kedua skenario tersebut memperoleh *accuracy* yang sama, skenario dengan *split* data 80:20, *epoch* maksimum 80, dan *batch size* 16 menghasilkan *test loss* lebih rendah, yaitu sebesar 0,0852 dibandingkan skenario *split* data 80:20, *epoch* maksimum 50, dan *batch size* 32 yang memperoleh *test loss* sebesar 0,0924. Oleh karena itu, skenario *split* data 80:20, *epoch* maksimum 80, dan *batch size* 16 dipilih sebagai skenario terbaik. Konfigurasi pemodelan CNN-LSTM pada skenario terbaik ditunjukkan pada Tabel 3.

Building of Informatics, Technology and Science (BITS)
Volume 8, No 1, June 2026 Page: 318−329
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v8i1.9917

Tabel 3. Konfigurasi Pemodelan CNN–LSTM pada Skenario Terbaik

Parameter	Nilai
Model	CNN–LSTM
CNN *Feature Extractor*	MobileNetV2
Ukuran *input* citra	224×224 piksel
Split data	80:20
Batch size	16
Epoch maximum	80
Learning rate awal	1e – 4
Learning rate fine-tuning	1e – 5
LSTM *units*	64
LSTM *timestep*	7
Fitur per *timestep*	128
L2 *regularization*	1e – 4
Dropout	0.5
Maximum frames per folder video	15
Accuracy	0,9916
Test loss	0,0852

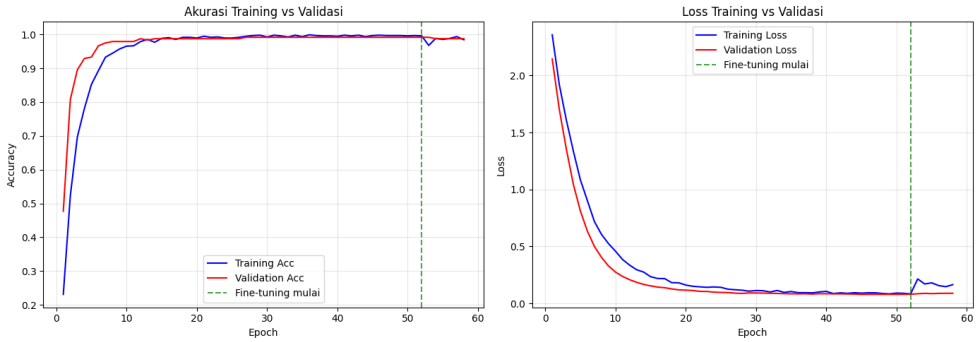
Pada konfigurasi terbaik, model menggunakan *split* data 80:20, *epoch* maksimum 80, dan *batch size* 16. Ukuran *input* citra yang digunakan adalah 224×224 piksel. Hasil ekstraksi fitur dari CNN kemudian disusun dalam bentuk sekuensial untuk diproses oleh LSTM. Model menggunakan 64 unit LSTM, 7 *timestep*, dan 128 fitur pada setiap *timestep*. Untuk mengurangi risiko *overfitting*, digunakan dropout sebesar 0.5 dan L2 regularization sebesar 1e-4. Selain itu, jumlah *frame* dibatasi maksimal 15 *frame* pada setiap folder video agar data dari video tertentu tidak mendominasi proses pelatihan.

Hasil evaluasi pada *test* set menunjukkan bahwa konfigurasi terbaik memperoleh *test accuracy* sebesar 0,9916 atau sekitar 99,2% dan *test loss* sebesar 0,0852. Hasil ini menunjukkan bahwa model mampu melakukan klasifikasi bangunan Melayu dan non-Melayu dengan tingkat kesalahan yang rendah. Selain *accuracy* dan *test loss*, evaluasi juga dilakukan menggunakan *classification report* untuk mengetahui nilai *precision, recall*, dan *F1-score*. Ringkasan hasil *precision, recall,* dan *F1-score* ditunjukkan pada Tabel 4.

Tabel 4. Ringkasan *Precision, Recall,* dan *F1-Score* pada Skenario Terbaik

Evaluasi	*Precision*	*Recall*	*F1-Score*
Macro Average	0,99	0,99	0,99
Weighted Average	0,99	0,99	0,99

Berdasarkan Tabel 4, nilai *macro average* dan *weighted average* untuk *precision, recall,* dan *F1-score* sebesar 0,99. Hasil tersebut menunjukkan bahwa model CNN-LSTM memiliki performa klasifikasi yang sangat baik pada data uji. Meskipun nilai akurasi yang diperoleh tinggi, hasil tersebut tetap ditafsirkan secara proporsional. Akurasi tidak dianalisis secara tunggal, tetapi disertai dengan *test loss, classification report, confusion matrix*, dan validasi GroupKFold. Selain itu, pembagian data telah memperhatikan kelompok sumber video untuk mengurangi risiko data *leakage*, sehingga evaluasi diarahkan pada kemampuan model mengenali karakteristik visual bangunan, bukan hanya kondisi perekaman tertentu. Grafik *accuracy* dan *loss training*-validasi pada Gambar 2 digunakan untuk melihat perkembangan performa model selama pelatihan.



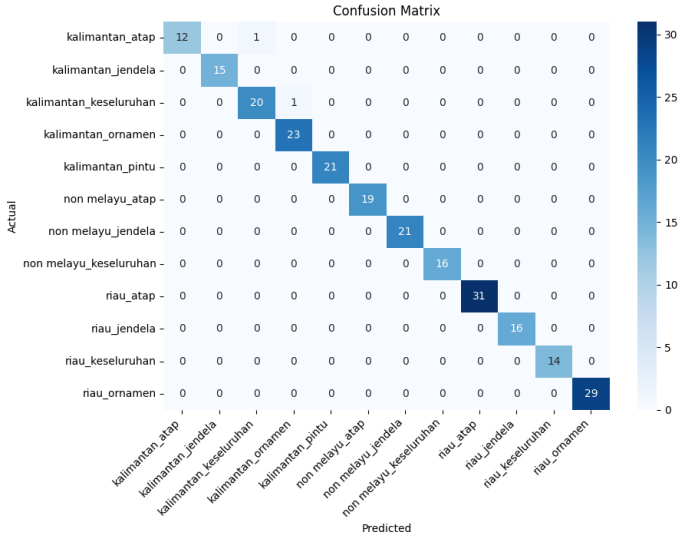
Gambar 2. Grafik *Accuracy* dan *Loss Training*-Validasi Model CNN–LSTM

Berdasarkan Gambar 2, nilai *training accuracy* dan *validation accuracy* meningkat secara signifikan pada *epoch* awal, kemudian cenderung stabil mendekati nilai 1.00. Hal ini menunjukkan bahwa model mampu mempelajari pola data dengan baik. Pada grafik *loss*, nilai *training loss* dan *validation loss* mengalami penurunan tajam pada awal pelatihan, kemudian stabil pada nilai rendah. Setelah tahap *fine-tuning* dimulai, terdapat sedikit fluktuasi pada *training*

Building of Informatics, Technology and Science (BITS)
Volume 8, No 1, June 2026 Page: 318−329
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v8i1.9917

loss, tetapi *validation loss* tetap rendah. Kondisi ini menunjukkan bahwa proses pelatihan berlangsung stabil pada skenario pengujian yang digunakan.

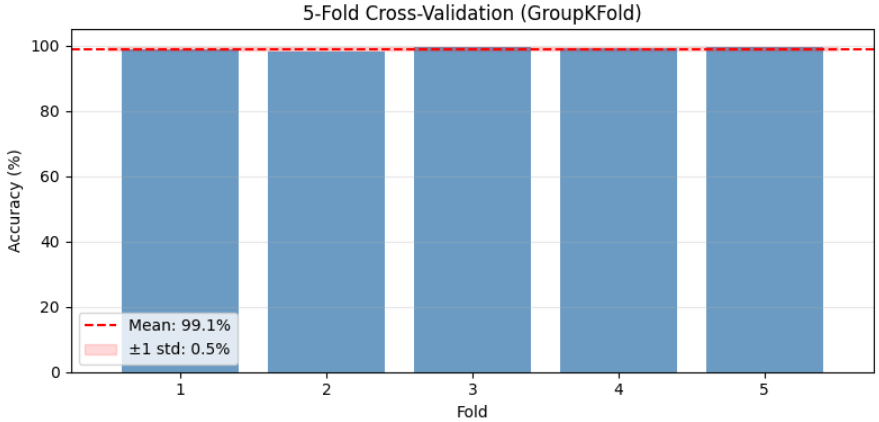
Untuk melihat pola kesalahan klasifikasi secara lebih rinci, digunakan *confusion matrix* pada skenario terbaik. Hasil *confusion matrix* model CNN–LSTM ditunjukkan pada Gambar 3.



Gambar 3. *Confusion Matrix* Model CNN–LSTM pada *Split* 80:20

Berdasarkan Gambar 3, sebagian besar data berada pada diagonal utama, yang menunjukkan bahwa model berhasil mengklasifikasikan mayoritas data sesuai kelas sebenarnya. Kesalahan klasifikasi hanya terjadi pada dua data, yaitu 1 data kelas kalimantan_atap yang diprediksi sebagai kalimantan_keseluruhan, serta 1 data kelas kalimantan_keseluruhan yang diprediksi sebagai kalimantan_ornamen. Kesalahan pada kelas kalimantan_atap dapat terjadi karena pada beberapa *frame*, bagian atap masih memuat konteks struktur bangunan lain, seperti dinding, tiang, bukaan, atau proporsi fasad, sehingga fitur visualnya berdekatan dengan tampilan keseluruhan bangunan. Sementara itu, kesalahan pada kelas kalimantan_keseluruhan yang diprediksi sebagai kalimantan_ornamen dapat terjadi karena elemen ornamen, ukiran, atau pola dekoratif pada fasad tampil dominan pada *frame* tertentu. Hal ini menunjukkan bahwa kesalahan klasifikasi terjadi pada kelas yang memiliki tumpang tindih fitur arsitektural, terutama antara atap, ornamen, dan tampilan keseluruhan bangunan.

Selain evaluasi menggunakan *test set*, dilakukan validasi tambahan menggunakan *K-Fold Cross-Validation* dengan metode GroupKFold sebanyak 5 *fold*. Validasi ini bertujuan untuk mengetahui kestabilan performa model pada beberapa pembagian data yang berbeda serta mengurangi risiko bias akibat pemilihan data tertentu. Penggunaan GroupKFold dilakukan agar data yang berasal dari kelompok atau sumber video yang sama tidak tersebar secara bersamaan pada data latih dan data uji, sehingga risiko data *leakage* dapat dikurangi. Hasil *K-Fold Cross-Validation* divisualisasikan pada Gambar 4.



Gambar 4. Grafik 5-*Fold Cross-Validation* menggunakan GroupKFold

Berdasarkan Gambar 4, *accuracy* pada setiap *fold* berada pada rentang 98,3% hingga 99,6%, dengan rata-rata *accuracy* sebesar 99,1% dan standar deviasi 0,5%. Nilai standar deviasi yang rendah menunjukkan bahwa performa model relatif konsisten pada beberapa pembagian data. Penggunaan GroupKFold juga memperkuat evaluasi karena pembagian data dilakukan dengan mempertimbangkan kelompok sumber video, sehingga *frame* dari sumber yang

Building of Informatics, Technology and Science (BITS)
Volume 8, No 1, June 2026 Page: 318−329
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v8i1.9917

sama tidak tersebar secara tidak terkendali pada data latih dan data uji. Dengan demikian, hasil evaluasi tidak hanya bergantung pada satu skenario *split* tertentu, tetapi juga menunjukkan kestabilan model pada beberapa pembagian data.

3.3 Analisis Hasil

Berdasarkan hasil evaluasi, model CNN-LSTM menunjukkan performa yang sangat baik dalam mengklasifikasikan bangunan Melayu dan non-Melayu. Pengujian dilakukan menggunakan beberapa skenario berdasarkan variasi *split* data, *epoch* maksimum, dan *batch size*. *Split* data yang digunakan terdiri dari rasio 70:30 dan 80:20. Dari seluruh skenario pengujian, nilai *accuracy* tertinggi sebesar 0.9916 diperoleh pada dua skenario, yaitu *split* data 80:20 dengan epoch maksimum 50 dan *batch size* 32, serta *split* data 80:20 dengan *epoch* maksimum 80 dan *batch size* 16.

Pemilihan skenario terbaik tidak hanya didasarkan pada nilai *accuracy*, tetapi juga mempertimbangkan nilai *test loss*. Skenario dengan *split* data 80:20, *epoch* maksimum 80, dan *batch size* 16 memperoleh *test loss* sebesar 0,0852, lebih rendah dibandingkan dengan skenario *split* data 80:20, *epoch* maksimum 50, dan *batch size* 32 yang memperoleh *test loss* sebesar 0,0924. Dengan demikian, skenario *split* data 80:20, *epoch* maksimum 80, dan *batch size* 16 dipilih sebagai konfigurasi terbaik karena menghasilkan *accuracy* tertinggi dengan nilai *loss* yang lebih rendah.

Hasil ringkasan *classification report* menunjukkan bahwa nilai *macro average* dan *weighted average* pada *precision, recall*, dan *F1-score* masing-masing mencapai 0,99. Nilai tersebut menunjukkan bahwa model memiliki performa klasifikasi yang tinggi baik secara rata-rata antarkelas maupun berdasarkan distribusi jumlah data pada setiap kelas. Hal ini mengindikasikan bahwa model CNN–LSTM mampu mempertahankan kemampuan klasifikasi yang baik pada data uji.

Hasil grafik *accuracy* dan *loss training-validasi* menunjukkan bahwa proses pelatihan berlangsung stabil. Nilai *training accuracy* dan *validation accuracy* meningkat pada *epoch* awal dan kemudian stabil pada nilai tinggi. Sementara itu, *training loss* dan *validation loss* mengalami penurunan pada awal pelatihan dan cenderung stabil pada nilai rendah. Setelah tahap *fine-tuning*, terdapat sedikit fluktuasi pada *training loss*, tetapi *validation loss* tetap rendah. Kondisi ini menunjukkan bahwa proses *fine-tuning* tidak menyebabkan *overfitting* yang signifikan dan model tetap memiliki kemampuan generalisasi yang baik.

Berdasarkan *confusion matrix*, kesalahan klasifikasi yang terjadi masih dapat dijelaskan secara arsitektural. Kelas kalimantan_atap dapat terprediksi sebagai kalimantan_keseluruhan karena bagian atap pada beberapa *frame* masih memuat konteks struktur bangunan lain. Sementara itu, kelas kalimantan_keseluruhan dapat terprediksi sebagai kalimantan_ornamen karena elemen ornamen atau detail fasad tampil cukup dominan. Dengan demikian, kesalahan klasifikasi tidak terjadi secara acak, tetapi muncul pada kelas yang memiliki kedekatan fitur visual.

Validasi tambahan menggunakan GroupKFold memperkuat hasil evaluasi karena model memperoleh rata-rata *accuracy* sebesar 99,1% dengan standar deviasi 0,5%. Nilai tersebut menunjukkan bahwa performa model relatif stabil pada beberapa pembagian data dan tidak hanya bergantung pada satu skenario pengujian tertentu.

Secara umum, performa model dipengaruhi oleh kombinasi CNN dan LSTM yang memiliki fungsi saling melengkapi. CNN berperan mengekstraksi fitur visual dari setiap *frame*, sedangkan LSTM mempelajari keterkaitan visual antarframe. Dengan demikian, model tidak hanya memanfaatkan informasi pada satu *frame*, tetapi juga mempertimbangkan rangkaian informasi visual dari video.

Perbandingan dengan penelitian terdahulu diarahkan pada domain arsitektur, bangunan, dan warisan budaya agar relevan secara substansi. Siountri dan Anagnostopoulos menerapkan metode *deep learning* berbasis YOLO untuk klasifikasi bangunan warisan budaya di Athena. Penelitian tersebut memperoleh tingkat keberhasilan 92% pada kelas bangunan neoklasik, 77,8% pada bangunan interwar, 71% pada *apartment building*, serta rata-rata 83,5% berdasarkan klasifikasi periode konstruksi. Pada pengujian data baru, penelitian tersebut berhasil mengklasifikasikan 31 dari 41 bangunan dengan tingkat keberhasilan 75,6% dan 91 dari 120 foto uji dengan tingkat keberhasilan 75,8% [30]. Penelitian lain oleh Tasir dan Khalid mengembangkan model HistoNet untuk klasifikasi citra tempat bersejarah. Model tersebut menggabungkan CNN, Transformer, dan Mamba *state-space* model untuk mengenali elemen lokal arsitektur serta konteks visual global. Hasil penelitian menunjukkan bahwa HistoNet memperoleh *accuracy* sebesar 95,66% dengan F1-*score* 0,943 pada dataset elemen warisan arsitektur, serta *accuracy* sebesar 96,46% pada dataset pengenalan bangunan bersejarah [9]. Sementara itu, Pramono *et al*. menerapkan Vision Transformer untuk mengidentifikasi gaya arsitektur bangunan Melayu, khususnya Rumah Melayu Riau dan Rumah Tradisional Melayu Pontianak. Penelitian tersebut memperoleh *precision* dan *recall* sebesar 0,99 pada data latih, serta nilai 0,98 untuk Rumah Melayu Riau dan 0,97 untuk Rumah Tradisional Melayu Pontianak pada data uji [31].

Dibandingkan dengan penelitian-penelitian tersebut, penelitian ini memperoleh *accuracy* sebesar 0,9916, *test loss* sebesar 0,0852, nilai *macro average* dan *weighted average F1-score* sebesar 0,99, serta rata-rata *accuracy* GroupKFold sebesar 99,1% dengan standar deviasi 0,5%. Hasil tersebut menunjukkan bahwa model CNN-LSTM memiliki performa yang kompetitif pada identifikasi bangunan Melayu dan non-melayu berbasis *frame* video. Namun, perbandingan performa tidak dapat dimaknai secara langsung karena setiap penelitian menggunakan dataset, jumlah kelas, objek, metode, dan skenario evaluasi yang berbeda. Oleh karena itu, hasil penelitian ini lebih tepat dipahami sebagai performa terbaik model pada dataset dan skenario evaluasi yang digunakan, bukan sebagai klaim bahwa model unggul secara mutlak pada seluruh kondisi data bangunan Melayu.

Secara keseluruhan, hasil penelitian menunjukkan bahwa model CNN-LSTM mampu memberikan performa yang sangat baik dalam mengklasifikasikan bangunan Melayu dan non-Melayu. Konfigurasi terbaik diperoleh pada

Building of Informatics, Technology and Science (BITS)
Volume 8, No 1, June 2026 Page: 318−329
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v8i1.9917

split data 80:20, *epoch* maksimum 80, dan *batch size* 16 dengan *accuracy* sebesar 0.9916 dan *test loss* sebesar 0.0852. Hasil ringkasan *classification report* dengan nilai *macro average* dan *weighted average F1-score* sebesar 0,99, serta *validasi* GroupKFold dengan rata-rata *accuracy* 99,1% dan standar deviasi 0,5%, menunjukkan bahwa model memiliki performa klasifikasi yang tinggi dan stabil pada beberapa pembagian data.

3.4 Implementasi Sistem

Model CNN-LSTM yang telah dilatih menggunakan konfigurasi terbaik diimplementasikan ke dalam aplikasi berbasis *web* menggunakan *framework* Flask. Sistem ini dirancang agar pengguna dapat mengunggah file berupa gambar maupun video bangunan, kemudian sistem akan memproses *input* tersebut dan menampilkan hasil identifikasi berdasarkan model yang telah dibangun. Tampilan antarmuka aplikasi ditunjukkan pada Gambar 5.

Gambar 5. Tampilan Antarmuka Aplikasi Identifikasi Bangunan Melayu

Berdasarkan Gambar 5, aplikasi menyediakan area unggah file dan tombol Analisis Sekarang untuk memulai proses identifikasi. Antarmuka dibuat sederhana agar pengguna dapat menggunakan sistem secara mudah dan interaktif. Pada *input* gambar, sistem melakukan *preprocessing* berupa pembacaan citra, perubahan ukuran menjadi 224×224 piksel, normalisasi, dan prediksi kelas menggunakan model. Pada *input* video, sistem terlebih dahulu mengekstraksi video menjadi beberapa *frame*, kemudian setiap *frame* diproses melalui tahapan yang sama seperti *input* gambar. Hasil prediksi dari beberapa *frame* digunakan untuk menentukan kelas akhir video.

Hasil prediksi ditampilkan dalam bentuk label kelas, kategori bangunan, nilai *confidence*, serta detail hasil prediksi. Dengan adanya kategori non-Melayu, sistem dapat memberikan indikasi apabila *input* yang diuji bukan termasuk bangunan Melayu. Secara keseluruhan, implementasi ini menunjukkan bahwa model CNN-LSTM dapat diterapkan dalam bentuk prototipe aplikasi *web* untuk membantu mengidentifikasi bangunan Melayu dan non-melayu secara lebih cepat dan interaktif.

4. KESIMPULAN

Penelitian ini menunjukkan bahwa model CNN-LSTM mampu mengidentifikasi bangunan Melayu dan non-Melayu berbasis data video yang diekstraksi menjadi *frame* dengan performa yang tinggi dan stabil. Konfigurasi terbaik diperoleh pada split data 80:20, *epoch* maksimum 80, dan *batch size* 16 dengan *accuracy* sebesar 0,9916, *test loss* sebesar 0,0852, nilai *macro average* dan *weighted average F1-score* sebesar 0,99, serta rata-rata *accuracy* GroupKFold sebesar 99,1% dengan standar deviasi 0,5%. Secara arsitektural, hasil tersebut menunjukkan bahwa model mampu menangkap karakteristik visual bangunan melalui elemen bentuk atap, jendela, pintu, ornamen, komposisi fasad, dan tampilan keseluruhan bangunan. Kesalahan klasifikasi yang terjadi pada kelas kalimantan_atap, kalimantan_keseluruhan, dan kalimantan_ornamen menunjukkan adanya kedekatan fitur visual antarbagian bangunan, terutama ketika atap masih memuat konteks struktur lain atau ketika ornamen tampil dominan pada fasad. Meskipun demikian, sistem masih memiliki keterbatasan pada kondisi lapangan tertentu, seperti bangunan yang tertutup pohon, kendaraan, pagar, atau objek lain, serta kondisi pencahayaan ekstrem, sudut pengambilan gambar yang terlalu miring, jarak kamera yang jauh, dan kualitas video yang rendah. Oleh karena itu, penelitian selanjutnya disarankan untuk memperluas dataset dengan variasi objek, lokasi, pencahayaan, tingkat oklusi, sudut kamera, dan kondisi bangunan yang lebih beragam agar kemampuan sistem dapat diuji secara lebih menyeluruh.

REFERENCES

[1] W. Huang and L. Chen, “Digital design of architectural heritage protection based on 3D reconstruction technology,” *Discov. Artif. Intell.*, vol. 5, p. Art. no. 248, 2025, doi: 10.1007/s44163-025-00504-5.
[2] D. Zhang, “Artificial Intelligence for the Preservation and Transmission of Non-Material Cultural Heritage: Opportunities, Ethical Challenges, and Future Directions,” *Appl. Comput. Eng.*, vol. 174, no. 1, pp. 214–220, 2025, doi: 10.54254/2755-2721/2025.PO25257.
[3] W. Zhang, “Reimagining cultural heritage conservation through VR , metaverse , and digital twins : An AI and blockchain-based framework,” *PLoS One*, vol. 20, no. 11, p. Art. no. e0335943, 2025, doi: 10.1371/journal.pone.0335943.
[4] J. Hutson, J. Weber, and A. Russo, “Digital Twins and Cultural Heritage Preservation : A Case Study of Best Practices and

Building of Informatics, Technology and Science (BITS)
Volume 8, No 1, June 2026 Page: 318−329
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v8i1.9917

Reproducibility in Chiesa dei SS Apostoli e Biagio," *Art Des. Rev.*, vol. 11, no. 1, pp. 15–41, 2023, doi: 10.4236/adr.2023.111003.
[5] F. Gîrbacia, "An Analysis of Research Trends for Using Artificial Intelligence in Cultural Heritage," *Electronics*, vol. 13, no. 18, p. Art. no. 3738, 2024, doi: 10.3390/electronics13183738.
[6] A. Heritage *et al.*, "Research on Image Classification and Retrieval Using Deep Learning with Attention Mechanism on Diaspora Chinese Architectural Heritage in Jiangmen, China," *Buildings*, vol. 13, no. 2, p. Art. no. 275, 2023, doi: 10.3390/buildings13020275.
[7] M. Wallace, V. Poulopoulos, and A. Antoniou, "An Overview of Big Data Analytics for Cultural Heritage," *Big Data Cogn. Comput.*, vol. 7, no. 1, p. Art. no. 14, 2023, doi: 10.3390/bdcc7010014.
[8] C. Frescoes, "The Application of ResNet-34 Model Integrating Transfer Learning in the Recognition and Classification of Overseas," *Electronics*, vol. 12, no. 17, p. Art. no. 3677, 2023, doi: 10.3390/ electronics12173677.
[9] A. M. Tasir, M. Nor, and A. Khalid, "OPEN A novel deep neural model for efficient and scalable historical place image classification," *Sci. Rep.*, vol. 15, p. Art. no. 42745, 2025, doi: 10.1038/s41598-025-26897-y 1.
[10] M. Mishra and P. B. Lourenço, "Artificial intelligence-assisted visual inspection for cultural heritage : State-of-the-art review," *J. Cult. Herit.*, vol. 66, pp. 536–550, 2024, doi: 10.1016/j.culher.2024.01.005.
[11] M. Alzahrani, M. Usman, S. K. Jarraya, S. Anwar, and T. Helmy, "Deep models for multi-view 3D object recognition: a review," *Artif. Intell. Rev.*, vol. 57, no. 12, p. Art. no. 323, 2024, doi: 10.1007/s10462-024-10941-w.
[12] W. Wang, X. Wang, G. Chen, and H. Zhou, "Multi-view SoftPool attention convolutional networks for 3D model classification," *Front. Neurorobot.*, vol. 16, p. Art. no. 1029968, 2022, doi: 10.3389/fnbot.2022.1029968.
[13] A. Pandey, P. Kumar, M. Manish, G. Ramraj, and G. Choudhary, "A Deep Learning-Based Hybrid CNN-LSTM Model for Location-Aware Web Service Recommendation," *Neural Process. Lett.*, vol. 56, no. 5, p. 248, 2024, doi: 10.1007/s11063-024-11687-w.
[14] M. Mao and A. Lee, "Deep Learning Innovations in Video Classification : A Survey on Techniques and Dataset Evaluations," *Electronics*, vol. 13, no. 14, pp. 1–30, 2024, doi: 10.3390/ electronics13142732.
[15] X. An, S. Li, and T. Wu, "Modeling Nonlinear Aeroelastic Forces for Bridge Decks with Various Leading Edges Using LSTM Networks," *Appl. Sci.*, vol. 13, no. 10, p. 6005, 2023, doi: 10.3390/app13106005.
[16] G. Gao *et al.*, "CNN-Bi-LSTM : A Complex Environment-Oriented Cattle Behavior Classification Network Based on the Fusion of CNN," *Sensors*, vol. 23, no. 18, p. 7714, 2023, doi: 10.3390/s23187714.
[17] R. V. Bidwe *et al.*, "Deep Learning Approaches for Video Compression : A Bibliometric Analysis," *Big Data Cogn. Comput.*, vol. 6, no. 2, p. 44, 2022, doi: 10.3390/ bdcc6020044.
[18] B. A. Khan and J. Jung, "Deep Learning-Based Human Activity Recognition Using Dilated CNN and LSTM on Video Sequences of Various Actions Dataset," *Appl. Sci.*, vol. 15, no. 22, p. 12173, 2025, doi: 10.3390/ app152212173.
[19] A. Rehman, S. B. Belhaouari, and A. Kabir, "On the Use of Deep Learning for Video Classification," *Appl. Sci.*, vol. 13, no. 12, p. Art. no. 2007, 2023, doi: 10.3390/ app13032007.
[20] S. Tipper, H. F. Atlam, and H. S. Lallie, "An Investigation into the Utilisation of CNN with LSTM for Video Deepfake Detection," *Appl. Sci.*, vol. 14, no. 21, p. no. 9754, 2024, doi: 10.3390/app14219754.
[21] J. Yoon and M. K. Choi, "Exploring Video Frame Redundancies for Efficient Data Sampling and Annotation in Instance Segmentation," *IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Work.*, vol. 2023-June, pp. 3308–3317, 2023, doi: 10.1109/CVPRW59228.2023.00333.
[22] R. J. Babić, "Cultural heritage image classification using transfer learning for feature extraction: a comparison," *CEUR Workshop Proc.*, vol. 3266, 2022.[Online]. Available: https://ceur-ws.org/Vol-3266/paper4.pdf
[23] J. Rodziewicz, A. Mielcarek, W. Janczukowicz, K. Bryszewski, and A. Jabło, "Technological Parameters of Rotating Electrochemical and Electrobiological Disk Contactors Depending on the Effluent Quality Requirements," *Appl. Sci.*, vol. 12, no. 11, p. 5503, 2022, doi: 10.3390/app12115503.
[24] A. M. Halim, M. D. P, and F. Nhita, "Handling Imbalanced Data Sets Using SMOTE and ADASYN to Improve Classification Performance of Ecoli Data Sets," *Build. Informatics, Technol. Sci.*, vol. 5, no. 1, pp. 246–253, 2023, doi: 10.47065/bits.v5i1.3647.
[25] A. Bensaoud and J. Kalita, "CNN-LSTM and Transfer Learning Models for Malware Classification based on Opcodes and API Calls," *Knowledge-Based Syst.*, 2023, doi: 10.1016/j.knosys.2024.111543.
[26] D. Akinola, A. O. Oyedemi, and M. O. Ajinaja, "A Deep Learning - based Hybrid CNN - LSTM Model for Human Activity Recognition," *Int. J. Sci. Res. Comput. Sci. Eng.*, vol. 12, no. 6, pp. 66–74, 2024.
[27] L. N. I. Afida, F. A. Bachtiar, and I. Cholissodin, "Klasifikasi Aktivitas Manusia Menggunakan Metode Long Short-Term Memory," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 11, no. 2, pp. 357–368, 2024, doi: 10.25126/jtiik.20241127060.
[28] B. Septian, C. Putra, and I. Tahyudin, "JITE (Journal of Informatics and Telecommunication Engineering) Performance Evaluation of CNN-LSTM and CNN-FNN Combinations for," *JITE (Journal Informatics Telecommun. Eng.*, vol. 8, no. 2, pp. 196–207, 2025, doi: 10.31289/jite.v8i2.13503.
[29] A. Nur, H. Mubarok, E. Nur, F. Dewi, and R. Edwinda, "Implementation of Convolutional Neural Network and Long Short-Term Memory Algorithms in Human Activity Recognition Based on Visual Processing Video," *JOIV Int. J. Informatics Vis.*, vol. 7, no. June, pp. 494–501, 2023, doi: 10.30630/joiv.7.2.1504.
[30] K. Siountri and C. N. Anagnostopoulos, "The Classification of Cultural Heritage Buildings in Athens Using Deep Learning Techniques," *Heritage*, vol. 6, no. 4, pp. 3673–3705, 2023, doi: 10.3390/heritage6040195.
[31] H. Pramono, S. Winiarti, and A. Fadlil, "Identifying Traditional Malay Building Architectural Styles Using Vision Transformer Architecture," *Int. J. Informatics Comput.*, vol. 7, no. 2, pp. 672–689, 2025, doi: 10.35842/ijicom.