INTEGRATION OF BERT-VAD, MFCC-DELTA, AND VGG16 IN TRANSFORMER-BASED FUSION ARCHITECTURE FOR MULTIMODAL EMOTION CLASSIFICATION

LAPORAN NON-REGULER

Diajukan untuk memenuhi salah satu syarat mencapai derajat Sarjana
Program Studi Informatika

disusun oleh
FISAN SYAFA NAYOMA
22.11.4836

FAKULTAS ILMU KOMPUTER
UNIVERSITAS AMIKOM YOGYAKARTA
YOGYAKARTA
2025

INTEGRATION OF BERT-VAD, MFCC-DELTA, AND VGG16 IN TRANSFORMER-BASED FUSION ARCHITECTURE FOR MULTIMODAL EMOTION CLASSIFICATION

LAPORAN NON-REGULER

Diajukan untuk memenuhi salah satu syarat mencapai derajat Sarjana
Program Studi Informatika

disusun oleh

FISAN SYAFA NAYOMA

22.11.4836

FAKULTAS ILMU KOMPUTER

UNIVERSITAS AMIKOM YOGYAKARTA

YOGYAKARTA

2025

HALAMAN PERSETUJUAN

JALUR NON-REGULER

INTEGRATION OF BERT-VAD, MFCC-DELTA, AND VGG16 IN TRANSFORMER-BASED FUSION ARCHITECTURE FOR MULTIMODAL EMOTION CLASSIFICATION

yang disusun dan diajukan oleh

Fisan Syafa Nayoma

22.11.4836

telah disetujui oleh Dosen Pembimbing
pada tanggal 27 Oktober 2025

Dosen Pembimbing,

Kusnawi, S.Kom., M.Eng.
NIK. 19030202112

HALAMAN PENGESAHAN

JALUR NON-REGULER

INTEGRATION OF BERT-VAD, MFCC-DELTA, AND VGG16 IN TRANSFORMER-BASED FUSION ARCHITECTURE FOR MULTIMODAL EMOTION CLASSIFICATION

yang disusun dan diajukan oleh

Fisan Syafa Nayoma

22.11.4836

Telah dipertahankan di depan Dewan Penguji
pada tanggal 27 Oktober 2025

Susunan Dewan Penguji

Nama Penguji	Tanda Tangan
Arifiyanto Hadinegoro, S.Kom., M.T. **NIK. 190302289**	
Bambang Pilu Hartato, S.Kom., M.Eng. **NIK. 190302707**	
Kusnawi, S.Kom., M.Eng. **NIK. 190302112**	

Laporan ini telah diterima sebagai salah satu persyaratan
untuk memperoleh gelar Sarjana Komputer
Tanggal 27 Oktober 2025

DEKAN FAKULTAS ILMU KOMPUTER

Prof. Dr. Kusrini, S.Kom., M.Kom.
NIK. 190302106

HALAMAN PERNYATAAN KEASLIAN KARYA

Yang bertandatangan di bawah ini,

Nama mahasiswa : Fisan Syafa Nayoma
NIM : 22.11.4836

Menyatakan bahwa laporan dengan judul berikut:

INTEGRATION OF BERT-VAD, MFCC-DELTA, AND VGG16 IN TRANSFORMER-BASED FUSION ARCHITECTURE FOR MULTIMODAL EMOTION CLASSIFICATION

Dosen Pembimbing: Kusnawi, S.Kom., M.Eng.

1. Karya tulis ini adalah benar-benar ASLI dan BELUM PERNAH diajukan untuk mendapatkan gelar akademik, baik di Universitas AMIKOM Yogyakarta maupun di Perguruan Tinggi lainnya.
2. Karya tulis ini merupakan gagasan, rumusan dan kegiatan SAYA sendiri, tanpa bantuan pihak lain kecuali arahan dari Dosen Pembimbing.
3. Dalam karya tulis ini tidak terdapat karya atau pendapat orang lain, kecuali secara tertulis dengan jelas dicantumkan sebagai acuan dalam naskah dengan disebutkan nama pengarang dan disebutkan dalam Daftar Pustaka pada karya tulis ini.
4. Perangkat lunak yang digunakan dalam penelitian ini sepenuhnya menjadi tanggung jawab SAYA, bukan tanggung jawab Universitas AMIKOM Yogyakarta.
5. Pernyataan ini SAYA buat dengan sesungguhnya, apabila di kemudian hari terdapat penyimpangan dan ketidak-benaran dalam pernyataan ini, maka SAYA bersedia menerima SANKSI AKADEMIK dengan pencabutan gelar yang sudah diperoleh, serta sanksi lainnya sesuai dengan norma yang berlaku di Perguruan Tinggi.

Yogyakarta, 27 Oktober 2025

Yang Menyatakan,

Fisan Syafa Nayoma

HALAMAN PERSEMBAHAN

Dengan penuh rasa syukur kepada Allah SWT, karya ini saya persembahkan kepada:

1. Kedua orang tua saya tercinta, atas doa, kasih saying, dan dukungan tanpa henti.
2. Dosen pembimbing dan seluruh dosen di Program Studi Sarjana Informatika, atas ilmu dan bimbingan yang sangat luar biasa dan berharga.
3. Teman-teman seperjuangan, yang telah memberikan saran, masukan, dan kebersamaan selama proses penelitian ini.

KATA PENGANTAR

Puji syukur penulis panjatkan ke hadirat Allah SWT karena berkat rahmat, taufik, dan hidayah-Nya, penulis dapat menyelesaikan karya tulis dengan judul "Integration of BERT-VAD, MFCC-Delta, and VGG16 in Transformer-Based Fusion Architecture for Multimodal Emotion Classification" dengan baik. Penyusunan karya tulis ini tidak lepas dari bantuan, bimbingan, dan dukungan dari berbagai pihak. Oleh karena itu, penulis ingin menyampaikan ucapan terima kasih yang sebesar-besarnya kepada:

1. Bapak Prof. Dr. M. Suyanto, M.M. selaku Rektor Universitas AMIKOM Yogyakarta
2. Bapak Prof. Dr. Kusrini, S.Kom., M.Kom. selaku Dekan Fakultas Ilmu Komputer, Universitas AMIKOM Yogyakarta.
3. Ibu Eli Pujastuti, M.Kom. selaku Ketua Program Studi Informatika.
4. Bapak Kusnawi, S.Kom., M.Eng. selaku dosen pembimbing yang telah memberikan arahan, bimbingan, serta masukan yang berharga dalam proses penelitian dan penyusunan karya tulis ini.
5. Bapak Triyono dan Ibu Sumarni selaku orang tua penulis yang selalu memberikan kasih sayang, doa, serta dukungan moral dan material yang tidak pernah putus.

Penulis menyadari bahwa karya tulis ini masih jauh dari sempurna. Oleh karena itu, penulis mengharapkan saran dan kritik yang membangun demi perbaikan di masa mendatang. Semoga karya tulis ini dapat memberikan manfaat bagi pengembangan ilmu pengetahuan, khususnya di bidang Informatika.

Yogyakarta, 7 September 2025
Penulis

DAFTAR ISI

DAFTAR TABEL

DAFTAR GAMBAR

DAFTAR LAMBANG DAN SINGKATAN

Ω	Tahanan Listrik
μ	Konstanta gesekan
TP	True positive
TN	True negative
FP	False positive
FN	False negative
ANFIS	Adaptive Network Fuzzy Inference System
SVM	Support Vector Machines
BERT	Bidirectional Encoder Representations from Transformers
MFCC	Mel-Frequency Cepstral Coefficients
VGG16	Visual Geometry Group 16
NRC-VAD	National Research Council Valence-Arousal-Dominance Lexicon
k	Jumlah lipatan (fold) pada K-Fold Cross Validation
SMOTE	Synthetic Minority Over-sampling Technique
W_i, b_i	Bobot dan bias pada proyeksi fitur modalitas
ReLU	Rectified Linear Unit atau fungsi aktivassi
X_i	Representasi fitur modalitas ke-i
Q, K, V	Query, Key, Value pada mekanisme Attention
d_k	Key dimension atau dimensi key pada Attention
softmax	Fungsi aktivasi untuk normalisasi probabilitas
Concat	Operasi penggabungan
W^o	Bobot proyeksi linear setelah Multi-Head Attention
LayerNorm	Normalisasi layer pada Transformer
FFN	Feedforward Neural Network
Epoch	Jumlah perulangan penuh pada seluruh dataset
Dense	fully connected layer dalam Keras
num_heads	Jumlah kepala atensi
tf.stack	TensorFlow stact function
$\hat{y}$	Vektor keluaran prediksi probabilitas kelas

DAFTAR ISTILAH

Vektor	besaran yang mempunyai arah
Accuracy	Rasio prediksi benar dari keseluruhan data
Precision	Rasio prediksi benar positif dari seluruh prediksi positif
Recall	Rasio benar positif dari keseluruhan data positif
Valence	Dimensi afektif tingkat emosi bernuansa positif atau negatif
Arousal	Dimensi afektif tingkat intensitas emosi
Dominance	Dimensi afektif tingkat kendali dalam suatu emosi
Early fusion	Penggabungan modalitas pada tahap awal
Late fusion	Penggabungan modalitas pada tahap akhir
Transformer fusion	Penggabungan yang memanfaatkan arsitektur transformer

INTISARI

Emosi merupakan kondisi yang berperan penting dalam interaksi manusia dan menjadi fokus utama penelitian kecerdasan dalam memanfaatkan multimoda. Penelitian sebelumnya telah mengklasifikasikan emosi multimoda tetapi masih kurang optimal karena tidak mempertimbangkan kompleksitas emosi manusia secara keseluruhan. Meskipun menggunakan data multimoda, pemilihan ekstraksi fitur dan proses penggabungan masih kurang relevan untuk meningkatkan akurasi. Penelitian ini mencoba mengkategorikan emosi dan meningkatkan presisi melalui metodologi multimoda yang memanfaatkan Fusion berbasis Transformer. Data yang digunakan terdiri dari sintesis tiga modalitas: teks (diekstraksi melalui BERT dan dinilai melalui dimensi afektif NRC Valence, Arousal, dan Dominance), audio (diekstraksi melalui MFCC dan delta-delta2 dari dataset RAVDESS dan TESS), dan gambar (diekstraksi melalui VGG16 pada dataset FER-2013). Model dibangun dengan memetakan setiap fitur ke dalam representasi dimensi yang identik dan diproses melalui blok Transformer untuk mensimulasikan interaksi antar modalitas, yang dikenal sebagai interaksi tingkat fitur. Prosedur klasifikasi dijalankan melalui lapisan padat dengan aktivasi softmax. Evaluasi model dilakukan menggunakan Stratified K-Fold Cross Validation dengan k=10. Hasil evaluasi menunjukkan bahwa model mencapai akurasi 95% pada lipatan kesembilan. Hasil ini menunjukkan peningkatan signifikan dari penelitian sebelumnya pada tingkat fitur (73,55%), dan menggarisbawahi efektivitas kombinasi ekstraksi fitur dan Fusion berbasis Transformer. Penelitian ini berkontribusi pada bidang sistem yang sadar emosi dalam informatika, memfasilitasi interaksi yang lebih adaptif, empatik, dan cerdas antara manusia dan komputer dalam aplikasi praktis.

Kata kunci: BERT, MFCC, Emosi Multimoda, Penggabungan Transformer, VGG16.

ABSTRACT

Emotion is a condition that plays an important role in human interaction and is the main focus of intelligence research in utilizing multimodal. Previous studies have classified multimodal emotions but are still less than optimal because they do not consider the complexity of human emotions as a whole. Although using multimodal data, the selection of feature extraction and the merging process are still less relevant to improving accuracy. This study attempts to categorize emotions and improve precision through a multimodal methodology that utilizes Transformer-based Fusion. The data used consists of a synthesis of three modalities: text (extracted through BERT and assessed through the affective dimensions of NRC Valence, Arousal, and Dominance), audio (extracted through MFCC and delta-delta2 from the RAVDESS and TESS datasets), and images (extracted through VGG16 on the FER-2013 dataset). The model is built by mapping each feature into an identical dimensional representation and processed through a Transformer block to simulate the interaction between modalities, known as feature-level interactions. The classification procedure is run through a dense layer with softmax activation. Model evaluation was performed using Stratified K-Fold Cross Validation with k=10. The evaluation results showed that the model achieved 95% accuracy in the ninth fold. This result shows a significant improvement from previous research at the feature level (73.55%), and underlines the effectiveness of the combination of feature extraction and Transformer-based Fusion. This study contributes to the field of emotion-aware systems in informatics, facilitating more adaptive, empathetic, and intelligent interactions between humans and computers in practical applications.

Keyword: *BERT, MFCC, Multimodal Emotion, Transformer-Based Fusion, VGG16*