Analisis Akurasi dan Waktu Proses Deteksi Sentimen Menggunakan *Image Mel-Spectrogram*

Jutono Gondohanindijo*
Teknik dan Informatika, Universitas AKI, Semarang, Indonesia
E-mail : jutono.gondohanindijo@unaki.ac.id*
**Corresponding author*
Received 6 August 2025; Revised 29 August 2025; Accepted 31 August 2025

Abstrak - Dalam upaya meningkatkan interaksi manusia-mesin, penelitian deteksi sentimen sudah banyak dilakukan peneliti untuk tujuan tersebut. Seiring dengan berkembangnya Mesin Pembelajaran, penelitian ini akan membandingkan kemampuan empat model klasifikasi : CNN, CRNN, SVM, dan MLP—dalam mengidentifikasi sentimen berbasis gambar *Mel-spectrogram*. Penelitian ini memanfaatkan representasi *Mel-Spectrogram* dari 640 sampel image (gambar) spektrogram yang mencakup delapan kelompok kelas sentimen berbeda. Setelah melalui tahap praproses data gambar dan ekstraksi fitur, kinerja model dievaluasi menggunakan validasi silang 10-fold serta metrik akurasi, presisi, recall, dan F1-score. CNN dan CRNN mencapai akurasi tertinggi (100%), sedangkan SVM dan MLP mencapai 99,22%. Dari sisi waktu pelatihan, SVM membutuhkan waktu paking sedikit, yaitu sebesar 0,45 detik. Penelitian ini bertujuan untuk mengetahui efektivitas pendekatan image (gambar) *Mel-Spectrogram* dan menegaskan perlunya pertimbangan *trade-off* antara akurasi tinggi dan efisiensi komputasi dalam pemilihan model.

Kata Kunci – Analisis, *Mel-Spectogram*, Sentimen, Waktu Proses

Abstract - *In an effort to improve human-machine interaction, sentiment detection research has been widely conducted by researchers for this purpose. As Machine Learning evolves, this research will compare the ability of four classification models : CNN, CRNN, SVM, and MLP to identify image-based Mel-Spectrogram sentiment. This study utilizes Mel-Spectrogram representations from 640 spectrogram image samples, encompassing eight different sentiment class groups. After going thru the image data preprocessing and feature extraction stages, the model's performance was evaluated using 10-fold cross-validation and the accuracy, precision, recall, and F1-score metrics. CNN and CRNN achieved the highest accuracy (100%), while SVM and MLP reached 99.22%. In terms of training time, SVM required the least amount of time, at 0.45 seconds. This research aims to determine the effectiveness of the Mel-Spectrogram image approach and to emphasize the need to consider the trade-off between high accuracy and computational efficiency in model selection.*

Keywords – *Analysis, Mel-Spectogram, Sentiment, Processing Time*

1. PENDAHULUAN

Pengenalan sentimen secara otomatis telah muncul sebagai bidang penting dalam komputasi afektif, yang memungkinkan mesin untuk memahami kondisi afektif manusia melalui sinyal akustik. Kemampuan ini sangat berperan dalam meningkatkan interaksi manusia-komputer, khususnya dalam aplikasi seperti asisten virtual, pemantauan kesehatan mental, dan sistem dialog adaptif [1]. Tujuan utama dari Deteksi Sentimen adalah untuk menganalisis sinyal gambar, mengekstrak fitur-fitur yang relevan terhadap sentimen, dan mengklasifikasikannya ke dalam kategori sentimen yang telah ditentukan.

Metode konvensional sebagian besar bergantung pada fitur akustik buatan tangan seperti *Mel-Frequency Cepstral Coefficients* (MFCCs), *Linear Predictive Coding* (LPC), *pitch*, energi,

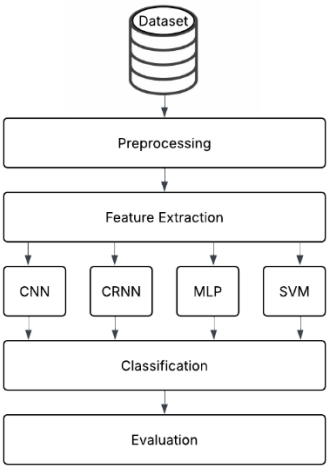
dan prosodi, yang kemudian dimasukkan ke dalam model klasifikasi tradisional seperti *Support Vector Machine* (SVM), *Naïve Bayes* (NB), atau *Random Forest* (RF) [2], [3]. Namun, pendekatan-pendekatan ini seringkali memiliki keterbatasan dalam menangkap kompleksitas temporal dan spasial dari sinyal audio, terutama saat menghadapi variasi sentimen yang halus. Kemajuan terbaru menunjukkan bahwa representasi visual dari audio seperti *Mel-Spectrogram* yang mengubah konten frekuensi jangka pendek menjadi gambar dua dimensi dapat secara signifikan meningkatkan akurasi deteksi sentimen. *Mel-Spectrogram* memanfaatkan sifat persepsi manusia dengan menerapkan skala logaritmik *Mel* pada pita frekuensi, sehingga cocok digunakan dalam model pembelajaran mendalam [4].

Penelitian ini mengeksplorasi performa dari empat model klasifikasi utama : *Convolutional Neural Networks* (CNN), *Convolutional Recurrent Neural Networks* (CRNN), *Support Vector Machines* (SVM), dan *Multi-Layer Perceptrons* (MLP) : dalam mengenali ekspresi berdasarkan representasi *Mel-Spectrogram*. CNN sangat efektif dalam mempelajari pola spasial dari gambar spektrogram, sementara CRNN mengintegrasikan fitur spasial dan temporal menggunakan unit rekuren seperti GRU atau LSTM, memberikan peningkatan dalam pemodelan sekuensial [5], [6]. Sementara itu, SVM dan MLP tetap dapat diandalkan dan efisien dalam skenario dengan dimensi yang telah direduksi menggunakan teknik seperti *Principal Component Analysis* (PCA) atau Standard Scaling [7], [8]. Sistem Deteksi Sentimen pada penelitian ini terdiri dari empat tahap utama: (1) *Mel-Spectrogram* dan transformasi gambar, (2) praproses dan normalisasi gambar (3) ekstraksi fitur dan pelatihan klasifikasi, dan (4) evaluasi performa menggunakan metrik standar seperti akurasi, presisi, recall, dan F1-score. Penelitian di bidang Deteksi Sentimen telah berkembang seiring kemajuan dalam teknik ekstraksi fitur dan model klasifikasi. Secara tradisional, para peneliti menggunakan fitur buatan tangan seperti MFCC, LPC, serta atribut prosodik seperti pitch, energi, dan durasi [9], [10], [11]. Fitur-fitur ini diketahui mencerminkan isyarat sentimen penting yang terdapat dalam sinyal audio.

Dalam beberapa studi terbaru, *Mel-Spectrogram* muncul sebagai representasi yang unggul karena mampu menangkap pola spektral dan temporal dari audio. Sebagai contoh, J. Li et al. [1] mengusulkan penggunaan spectrogram dua saluran untuk pengenalan sentimen berbasis CNN dan berhasil meningkatkan kinerja di berbagai kategori sentimen. Demikian pula, Colunga-Rodriguez et al. [12] mengembangkan dataset audio yang berfokus pada fitur audio sentimen yang diekstrak dari *spectrogram*. Untuk mengevaluasi performa model, dataset yang terdiri dari 640 sampel gambar berlabel delapan kelas sentimen telah dilakukan proses pengenalan pola dan evaluasi identifikasi kelasnya. Penelitian ini tidak hanya menilai akurasi klasifikasi, tetapi juga efisiensi waktu pelatihan, serta menyoroti pertukaran antara kompleksitas dan performa di antara model-model tersebut.

2. METODE PENELITIAN

Penelitian ini memfokuskan pada analisis performa akurasi dan waktu proses dari empat model klasifikasi : *Support Vector Machine* (SVM), *Convolutional Neural Network* (CNN), *Convolutional Recurrent Neural Network* (CRNN), dan *Multi-Layer Perceptron* (MLP) dalam mengklasifikasikan sentimen menggunakan *Mel-Spectrogram* sebagai fitur input utama. Gambar 1 menunjukkan Sistem yang diusulkan terdiri dari beberapa tahap: dataset gambar *Mel-spectrogram*, praproses gambar, ekstraksi fitur, klasifikasi menggunakan keempat model, dan evaluasi berdasarkan metrik seperti akurasi, presisi, *recall*, dan *F1-score*.



Gambar 1. Metodologi Deteksi Sntimen

A. *Support Vector Machine* (SVM)

SVM adalah algoritma pembelajaran mesin statistik yang digunakan untuk mencari hyperplane optimal untuk memisahkan data dari berbagai kelas dalam ruang yang sangat besar. Sebelum diklasifikasikan, spektrogram *Mel* diratakan menjadi vektor fitur dan diproses menggunakan PCA. Batas keputusan non-linier dapat ditangani dengan Fungsi Basis Radial (RBF). Karena kemampuan mereka untuk generalisasi pada dataset yang kecil dan tidak seimbang, SVM telah terbukti menghasilkan hasil yang dapat diandalkan dalam berbagai tugas pengenalan sentimen.

Dalam studi sebelumnya, SVM mencapai akurasi 100% di lingkungan terkendali dan 82,5% pada *Berlin Emotional Speech Database* [14]. Dalam eksperimen ini, SVM berperan sebagai baseline untuk mengevaluasi kekuatan relatif dari model pembelajaran mendalam yang memproses spectrogram sebagai gambar 2D.

B. *Convolutional Neural Network* (CNN)

CNN adalah model deep learning yang sangat cocok untuk tugas klasifikasi gambar. *Mel-spectrogram* dapat diperlakukan sebagai gambar grayscale, memungkinkan CNN untuk mengekstrak fitur hierarkis dari representasi waktu–frekuensi. Arsitektur CNN yang digunakan terdiri dari tiga lapisan konvolusi dengan aktivasi ReLU, diikuti dengan *max-pooling*, dan diakhiri dengan lapisan dense dengan aktivasi *softmax* untuk klasifikasi multi-kelas. Pendekatan ini telah digunakan dalam berbagai studi, dan terbukti memberikan performa yang tinggi dalam deteksi sentimen dari gambar [8].

Menurut Alluhaidan et al. [15], model berbasis CNN yang menggunakan representasi spektral seperti *Mel-spectrogram* dan fitur chroma dapat mencapai akurasi klasifikasi lebih dari 87%.

C. *Convolutional Recurrent Neural Network* (CRNN)

Untuk meningkatkan kemampuan pemodelan temporal dari CNN, arsitektur CRNN menggabungkan lapisan konvolusi dengan unit rekuren seperti GRU atau LSTM. CNN efektif untuk mengekstrak pola spasial pada spectrogram, sedangkan CRNN mampu menangkap ketergantungan sekuensial dari data audio.

Dalam implementasi ini, *Mel-spectrogram* diproses melalui lapisan CNN untuk ekstraksi fitur spasial, kemudian melalui GRU untuk menangkap dinamika temporal. Kombinasi

ini terbukti lebih unggul dibandingkan CNN saja, sebagaimana ditunjukkan oleh Begazo et al. [9].

D. *Multi-Layer Perceptron* (MLP)

MLP adalah jaringan saraf umpan-maju sepenuhnya yang bekerja pada fitur input berbentuk vektor. Dalam studi ini, gambar *Mel-spectrogram* diratakan dan dinormalisasi sebelum dimasukkan ke dalam model MLP. Arsitektur MLP terdiri dari beberapa lapisan tersembunyi dengan aktivasi ReLU, diakhiri dengan lapisan output softmax. Meskipun kurang mampu menangkap pola spasial dibanding CNN, MLP lebih ringan secara komputasi dan tetap memberikan pembandingan yang berguna.

Raghu dan Sadanandam [13] menunjukkan bahwa fitur spektral hibrida yang diproses dengan MLP mencapai akurasi hingga 81%.

3. HASIL DAN PEMBAHASAN

3.1 Dataset

Penelitian ini menggunakan 640 sampel gambar *Mel-Spectrogram* sentimen dalam format .png yang terdiri dari delapan kategori sentimen: cahaya jijik, cahaya senang, cahaya marah, cahaya sedih, dani marah, dani sedih, dani senang, dan dani jijik. Setiap kelas memiliki jumlah data yang seimbang.

3.2 Preprocessing

Praproses dilakukan dalam beberapa tahap untuk memproses data gambar *Mel-Spectrogram* menjadi data yang siap digunakan oleh model pembelajaran mesin.

1. *Mel-Spectrogram* menggunakan 128 Mel bands dan diubah ke skala desibel (dB).
2. Render Gambar Spektrogram: *Mel-Spectrogram* disimpan sebagai gambar grayscale dalam format .png.

3.3 Ekstraksi Fitur

1. Untuk CNN/CRNN, gambar disesuaikan menjadi ukuran 128×128 piksel.
2. Untuk SVM/MLP, gambar diratakan menjadi vektor 1D. MLP menggunakan StandardScaler, sedangkan SVM melalui proses PCA.

3.4 Klasifikasi

Proses klasifikasi menggunakan 4 model pengklasifikasi yaitu *Convolutional Neural Network* (CNN), *Convolutional Recurrent Neural Network* (CRNN), *Support Vector Machine* (SVM), dan *Multi-Layer Perceptron* (MLP)

3.5 Evaluasi

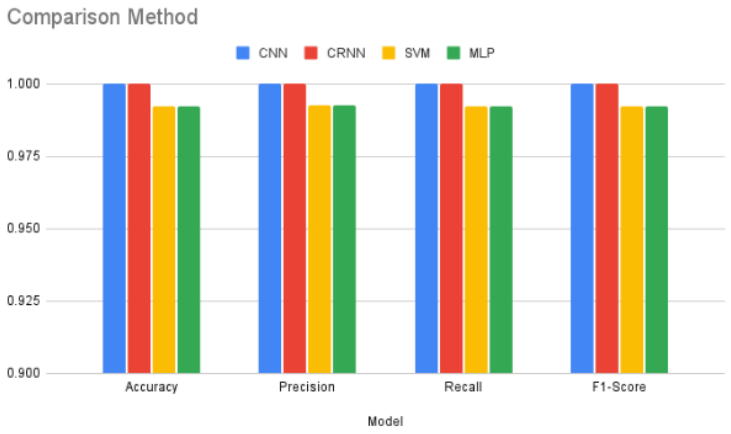
Semua model dievaluasi menggunakan 10-fold cross-validation. Hasilnya nampak pada Tabel 1 berikut:

Tabel 1. Hasil Perbandingan Evaluasi Model

Model	Accuracy	Precision	Recall	F1-Score	Training Time (s)
CNN	1.00	1.00	1.00	1.00	19.15
CRNN	1.00	1.00	1.00	1.00	19.06

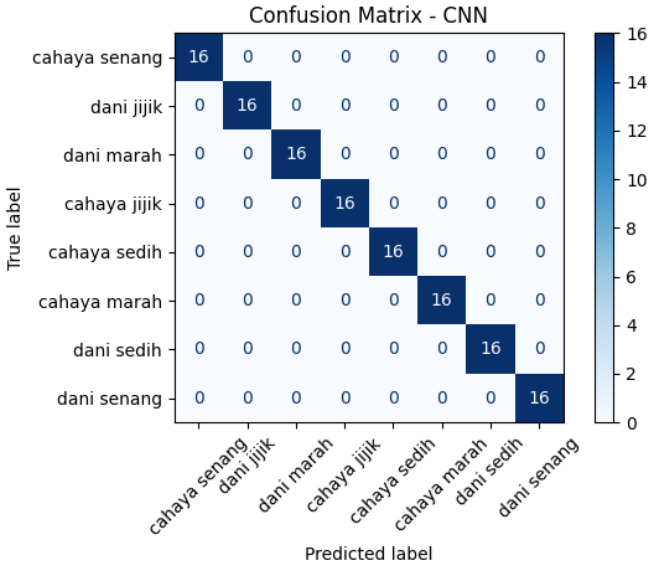
SVM	0.99	0.99	0.99	0.99	4.05
MLP	0.99	0.99	0.99	0.99	34.83

Tabel 1 menunjukkan hasil evaluasi performa dari empat model klasifikasi—CNN, CRNN, SVM, dan MLP—dalam tugas deteksi snetimen berbasis gambar *Mel-Spectrogram*. Model CNN dan CRNN berhasil mencapai akurasi sebesar 100%, dengan skor Presisi, Recall dan F1 sebesar 1.000 atau 100%, menunjukkan kemampuan yang tinggi untuk mengenali pola spasial dan temporal pada data gambar dalam bentuk *Mel-Spectogram*. Sementara itu, model SVM dan MLP mencapai akurasi sebesar 99,21%. Dalam hal efisiensi waktu pelatihan, SVM unggul dengan waktu hanya 4,05 detik, sedangkan MLP membutuhkan waktu paling lama, 34,83 detik. Hasil ini menunjukkan bahwa meskipun CNN dan CRNN memberikan akurasi maksimal, SVM tetap unggul dari sisi efisiensi komputasi, menjadikannya pilihan yang baik untuk lingkungan dengan keterbatasan sumber daya.



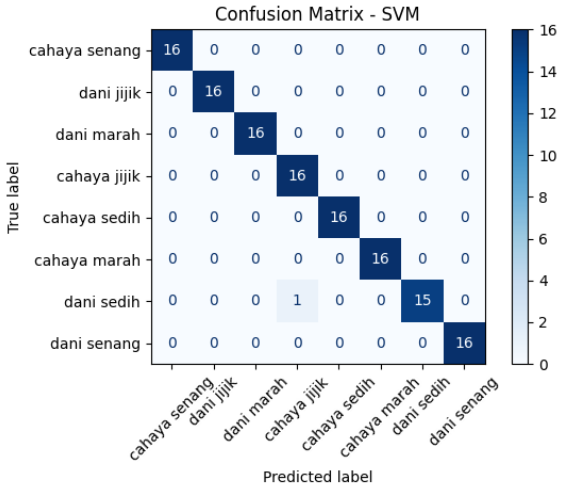
Gambar 2. Grafik Perbandingan Performa Klasifikasi

Grafik dalam Gambar 2 menunjukkan perbandingan bahwa CNN dan CRNN memiliki kinerja yang lebih unggul dibandingkan model tradisional dalam semua metrik evaluasi. Waktu pelatihan yang cepat pada SVM memberikan keuntungan dari segi efisiensi, sementara MLP membutuhkan waktu pelatihan yang lebih lama namun tetap menghasilkan performa yang kompetitif.



Gambar 3. *Confusion matrix* CNN

Gambar 4 menunjukkan bagaimana model *Convolutional Neural Network* (CNN) mengkategorikan delapan jenis sentimen: "cahaya senang", "cahaya jijik", "cahaya marah", "cahaya sedih", "dani sedih", "dani senang", "dani marah", dan "dani jijik". Setiap kelas memiliki 16 data uji, dan hasil klasifikasi menunjukkan bahwa model memprediksi semua data dengan benar. Nilai diagonal utama matriks, yang bernilai 16, ketika semua sel lainnya bernilai nol, menunjukkan hal ini. Oleh karena itu, model CNN memiliki akurasi sempurna sebesar 100%, yang menunjukkan bahwa data uji yang digunakan tidak mengandung kesalahan klasifikasi.



Gambar 4. *Confusion matrix* SVM

Gambar 5 menunjukkan *confusion matrix* SVM yang memperlihatkan hasil klasifikasi menggunakan model *Support Vector Machine* (SVM) pada dataset yang sama. Secara umum, model SVM menunjukkan performa yang sangat baik dengan memprediksi secara tepat seluruh kelas, kecuali pada satu kasus. Kesalahan terjadi pada kelas "dani sedih", di mana 15 dari 16 data diklasifikasikan dengan benar, sementara satu data keliru diprediksi sebagai "cahaya jijik". Meski demikian, seluruh kelas lainnya berhasil diprediksi dengan benar. Dengan hanya satu kesalahan

dari total 128 data uji, model SVM memperoleh tingkat akurasi sebesar 99,22%, yang mencerminkan kinerja klasifikasi yang sangat tinggi meskipun tidak sempurna.

4. KESIMPULAN

Studi ini membandingkan empat model mengenali sentimen menggunakan gambar *Mel-Spectrogram*: CNN, CRNN, SVM, dan MLP. Hasilnya menunjukkan bahwa CNN dan CRNN paling baik dengan akurasi 100%, sedangkan SVM dan MLP menghasilkan akurasi 99,22%. SVM juga paling cepat dalam hal kecepatan pelatihan (4,05 detik), sedangkan MLP paling lama (34,83 detik). Model CNN dan CRNN menghasilkan akurasi yang lebih baik meskipun membutuhkan waktu pelatihan lebih lama. Pilihan model bisa disesuaikan dengan tujuan dan kebutuhan aplikasi.

DAFTAR PUSTAKA

[1] Z. Fang, B. Yin, Z. Du, and X. Huang, "Fast environmental sound classification based on resource adaptive convolutional neural network," Sci. Rep., vol. 12, no. 1, pp. 1–18, 2022, doi: 10.1038/s41598-022-10382-x.

[2] T. Qiao, S. Zhang, S. Cao, and S. Xu, "High accurate environmental sound classification: Sub-spectrogram segmentation versus temporal-frequency attention mechanism," Sensors, vol. 21, no. 16, 2021, doi: 10.3390/s21165500.

[3] J. H. Chowdhury, S. Ramanna, and K. Kotecha, "Speech emotion recognition with light weight deep neural ensemble model using hand crafted features," Sci. Rep., vol. 15, no. 1, pp. 1–14, 2025, doi: 10.1038/s41598-025-95734-z.

[4] T. Liu, D. Yan, R. Wang, N. Yan, and G. Chen, "Identification of fake stereo audio using svm and cnn," Inf., vol. 12, no. 7, 2021, doi: 10.3390/info12070263.

[5] Y. K. Aini, T. B. Santoso, and T. Dutono, "Pemodelan CNN Untuk Deteksi Emosi Berbasis Speech Bahasa Indonesia," J. Komput. Terap., vol. 7, no. 1, pp. 143–152, 2021, doi: 10.35143/jkt.v7i1.4623.

[6] T. A. Ayall, C. Zhou, H. Liu, G. M. Brhanemeskel, S. T. Abate, and M. Adjeisah, "Amharic spoken digits recognition using convolutional neural network," J. Big Data, vol. 11, no. 1, 2024, doi: 10.1186/s40537-024-00910-z.

[7] K. A. Araño, P. Gloor, C. Orsenigo, and C. Vercellis, "When Old Meets New: Emotion Recognition from Speech Signals," Cognit. Comput., vol. 13, no. 3, pp. 771–783, 2021, doi: 10.1007/s12559-021-09865-2.

[8] L. Salam and R. S. Saxena, "Comparative Analysis of SVM and CNN for Hyperspectral Image Classification," pp. 4–7, 2024.

[9] R. Begazo, A. Aguilera, I. Dongo, and Y. Cardinale, "A Combined CNN Architecture for Speech Emotion Recognition," Sensors, vol. 24, no. 17, pp. 1–39, 2024, doi: 10.3390/s24175797.

[10] A. F. R. Nogueira, H. S. Oliveira, J. J. M. Machado, and J. M. R. S. Tavares, "Sound Classification and Processing of Urban Environments: A Systematic Literature Review," Sensors, vol. 22, no. 22, pp. 1–30, 2022, doi: 10.3390/s22228608.

[11] G. K. Chellamani, N. Aishwarya, C. Chandhana, K. Kaur, and R. T. S. Babu, "SpectroFusionNet a CNN approach utilizing spectrogram fusion for electric guitar play recognition," Sci. Rep., vol. 15, no. 1, pp. 1–19, 2025, doi: 10.1038/s41598-025-00287-w.

[12] A. A. Colunga-Rodriguez, A. Martínez-Rebollar, H. Estrada-Esquivel, E. Clemente, and O. A. Pliego-Martínez, "Developing a Dataset of Audio Features to Classify Emotions in Speech," Computation, vol. 13, no. 2, pp. 1–20, 2025, doi: 10.3390/computation13020039.

[13] K. Raghu and M. Sadanandam, "Emotion Recognition from Speech Utterances with Hybrid Spectral Features Using Machine Learning Algorithms," Trait. du Signal, vol. 39,

no. 2, pp. 603–609, 2022, doi: 10.18280/ts.390222.

[14] J. Li, X. Zhang, L. Huang, F. Li, S. Duan, and Y. Sun, "Speech Emotion Recognition Using a Dual-Channel Complementary Spectrogram and the CNN-SSAE Neutral Network," Appl. Sci., vol. 12, no. 19, 2022, doi: 10.3390/app12199518.

[15] A. S. Alluhaidan, O. Saidani, R. Jahangir, M. A. Nauman, and O. S. Neffati, "Speech Emotion Recognition through Hybrid Features and Convolutional Neural Network," Appl. Sci., vol. 13, no. 8, 2023, doi: 10.3390/app13084750.