

Analisis Kinerja CNN+LSTM dalam Klasifikasi Tangisan Bayi Menggunakan Ekstraksi Fitur Audio

^{1*}David Yoga Wicaksono, ²Daniel Swanjaya, ³Resty Wulanningrum

^{1,2,3} Teknik Informatika, Universitas Nusantara PGRI Kediri

¹davidyogawicaksono@gmail.com, ²daniel@unpkediri.ac.id, ³restyw@unpkdr.ac.id

Abstrak— Tangisan adalah alat komunikasi utama bagi bayi untuk menyampaikan kebutuhan, tetapi seringkali sulit dipahami. Studi ini bertujuan untuk mengklasifikasikan lima jenis tangisan bayi menggunakan ekstraksi fitur Mel Spectrogram. Untuk mengatasi ketidakseimbangan data, teknik augmentasi audio dan metode SMOTE diterapkan. Analisis perbandingan arsitektur dievaluasi dalam studi ini. Hasil pengujian menunjukkan bahwa model berbasis konvolusi memberikan kinerja tinggi di atas 95%. Meskipun Convolutional Neural Network memiliki akurasi sedikit lebih tinggi, model hibrida CNN-LSTM terbukti paling optimal karena menghasilkan nilai kesalahan pengujian terendah sebesar 0,1160. Integrasi CNN dan LSTM mampu menangkap fitur spasial-frekuensi serta karakteristik temporal suara secara komprehensif. Kesimpulannya, model hibrida ini memberikan solusi cerdas dan akurat untuk sistem pengenalan suara bayi guna mendukung pengembangan asisten pengasuh di masa depan.

Kata Kunci— CNN-LSTM, Deep Learning, Mel Spectrogram, SMOTE, Tangisan Bayi.

Abstract— Crying is a primary communication tool for infants to convey needs, but it is often difficult to understand. This study aims to classify five types of infant cries using Mel Spectrogram feature extraction. To address data imbalance, audio augmentation techniques and the SMOTE method are applied. A comparative analysis of architectures is evaluated in this study. Test results show that the convolution-based model provides high performance above 95%. Although the Convolutional Neural Network has slightly higher accuracy, the CNN-LSTM hybrid model proved to be the most optimal because it produced the lowest test error value of 0.1160. The integration of CNN and LSTM is able to comprehensively capture the spatial-frequency features as well as the temporal characteristics of sound. In conclusion, this hybrid model provides an intelligent and accurate solution for infant speech recognition systems to support the development of future caregiver assistants.

Keywords— Baby Cry, CNN-LSTM, Deep Learning, Mel Spectrogram, SMOTE

This is an open access article under the CC BY-SA License.



I. PENDAHULUAN

Tangisan bayi merupakan bentuk komunikasi utama selama periode neonatal [1],* namun interpretasinya seringkali ambigu karena kemiripan karakteristik akustik yang tinggi antara berbagai jenis tangisan [2]. Variasi frekuensi, intensitas, dan pola temporal yang serupa menyebabkan tingkat ambiguitas yang tinggi dalam proses interpretasi manual [3]. Kondisi ini menyulitkan manusia untuk membedakan berbagai kondisi bayi secara akurat, seperti lapar, sakit, atau ketidaknyamanan [4]. Interpretasi tangisan yang masih bergantung pada pengalaman individu berpotensi menyebabkan kesalahan identifikasi dan keterlambatan dalam memenuhi kebutuhan bayi [5]. Oleh karena itu, diperlukan pendekatan otomatis yang dapat menganalisis

karakteristik akustik tangisan bayi secara objektif dan konsisten untuk membantu proses identifikasi kondisi bayi dengan tingkat akurasi yang lebih tinggi.

Meskipun berbagai arsitektur Deep Learning telah menunjukkan kinerja yang baik dalam klasifikasi tangisan bayi [6], sebagian besar penelitian masih menggunakan satu jenis arsitektur secara independen sehingga kemampuan untuk menangkap karakteristik frekuensi dan temporal secara bersamaan masih terbatas [7]. Pendekatan yang hanya memanfaatkan CNN atau LSTM secara terpisah berpotensi kehilangan sebagian informasi penting yang terkandung dalam sinyal audio [8]. Oleh karena itu, diperlukan arsitektur hibrida CNN-LSTM yang dapat mengekstraksi fitur frekuensi melalui CNN dan mempelajari hubungan temporal melalui LSTM secara bersamaan [9]. Dengan kemampuan ini, CNN-LSTM diharapkan mampu meningkatkan kemampuan model dalam menangkap karakteristik akustik yang kompleks melalui integrasi informasi spasial dan temporal [10].

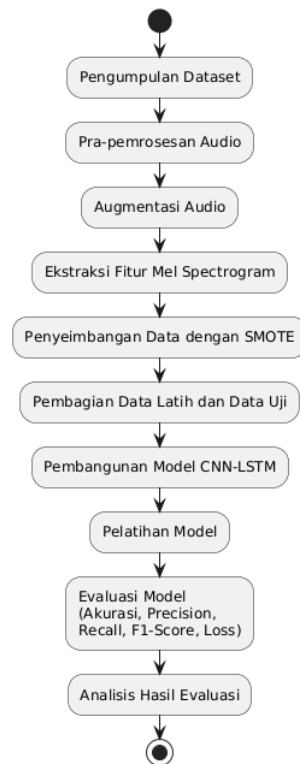
Berdasarkan permasalahan tersebut, penelitian ini bertujuan untuk mengembangkan model klasifikasi tangisan bayi menggunakan arsitektur hibrida Convolutional Neural Network (CNN) dan Long Short-Term Memory (LSTM) dengan fitur Mel Spectrogram sebagai representasi sinyal audio [11]. Penelitian ini berfokus pada analisis kinerja model CNN-LSTM dalam mengidentifikasi lima kategori tangisan bayi, yaitu lapar, sakit perut, ketidaknyamanan, sendawa, dan kelelahan. Evaluasi dilakukan menggunakan metrik akurasi, presisi, recall, dan F1-score untuk menilai kemampuan model dalam mengenali karakteristik akustik dari setiap kategori tangisan. Hasil penelitian ini diharapkan dapat berkontribusi pada pengembangan sistem identifikasi kondisi bayi berbasis kecerdasan buatan yang lebih akurat dan objektif.

II. METODE

A. Alur Penelitian

Penelitian ini dilakukan melalui beberapa tahapan sistematis yang dirancang untuk menghasilkan model klasifikasi yang akurat dan kuat. Tahapan tersebut meliputi pengumpulan dan pra-pemrosesan data audio, augmentasi audio, ekstraksi fitur menggunakan Mel Spectrogram, penanganan ketidakseimbangan data menggunakan metode SMOTE, pelatihan model menggunakan arsitektur hibrida CNN-LSTM [12], serta tahapan evaluasi performa model, pada gambar 1 merupakan alur penelitian yang akan dijalankan.

Alur Penelitian CNN-LSTM Klasifikasi Tangisan Bayi



Gambar 1. Alur Penelitian

B. Dataset dan Pra-pemrosesan Data

Dataset yang digunakan dalam penelitian ini berasal dari (github.com/gveres/donateacry-corpus) terdiri dari rekaman suara tangisan bayi yang terbagi ke dalam lima kategori kelas target, yaitu: lapar (hungry), nyeri (belly_pain), tidak nyaman (discomfort), bersendawa (burping), dan lelah (tired). Sebelum proses ekstraksi fitur, dilakukan augmentasi data audio untuk meningkatkan keragaman data pelatihan.

Teknik augmentasi yang diterapkan meliputi penambahan derau putih (white noise injection), pergeseran waktu (time shifting), dan perubahan nada (pitch shifting) [13]. Proses ini bertujuan meningkatkan kemampuan generalisasi model terhadap variasi kondisi perekaman. Setelah melalui proses augmentasi, fitur audio diekstraksi menggunakan representasi Mel Spectrogram melalui pustaka Librosa.

C. Ekstraksi Fitur Mel Spectrogram

Mel Spectrogram merupakan representasi waktu-frekuensi yang banyak digunakan dalam analisis sinyal audio karena mampu menggambarkan distribusi energi frekuensi sesuai karakteristik persepsi pendengaran manusia [14]. Parameter yang digunakan meliputi n_fft sebesar 2048, hop_length sebesar 512, dan jumlah filter Mel (n_mels) sebanyak 128.

Proses pembentukan Mel Spectrogram dimulai dengan penerapan Transformasi Fourier Waktu Singkat (STFT) untuk mengubah sinyal audio menjadi representasi frekuensi pada setiap segmen waktu. STFT dihitung menggunakan Persamaan (1):

$$X(m, k) = \sum_{n=0}^{N-1} x[n] w[n - mH] e^{\left\{ \frac{j2\pi kn}{N} \right\}} \quad (1)$$

Spektrum frekuensi yang diperoleh kemudian dipetakan ke dalam skala Mel menggunakan sekumpulan filter Mel untuk menyesuaikan representasi frekuensi dengan karakteristik pendengaran manusia. Konversi frekuensi dari satuan Hertz (Hz) ke skala Mel dilakukan menggunakan Persamaan (2):

$$M(f) = 2595 \times \log_{10} \left(1 + \frac{f}{700} \right) \quad (2)$$

Proses pembentukan Mel Spectrogram diawali dengan penerapan Short-Time Fourier Transform (STFT) untuk mengubah sinyal audio menjadi spektrum frekuensi pada setiap segmen waktu [15]. Energi spektral yang dihasilkan kemudian diproyeksikan ke dalam skala Mel menggunakan sekumpulan filter Mel sehingga menghasilkan matriks dua dimensi yang merepresentasikan karakteristik akustik sinyal secara lebih kompak dan informatif. Representasi ini selanjutnya digunakan sebagai masukan bagi model CNN dan CNN-LSTM dalam proses klasifikasi tangisan bayi.

D. Penanganan Ketidakseimbangan Data dengan SMOTE

Karakteristik utama dari dataset tangisan bayi dalam penelitian ini adalah adanya ketidakseimbangan jumlah sampel yang signifikan antar kelas (imbalanced dataset), di mana beberapa kelas seperti 'lapar' memiliki jumlah sampel yang jauh lebih mendominasi dibandingkan kelas 'bersendawa' atau 'lelah'. Jika dibiarkan, kondisi ini akan menyebabkan model mengalami bias dan cenderung memprioritaskan akurasi pada kelas mayoritas [16].

Untuk mengatasi masalah tersebut, metode Synthetic Minority Over-sampling Technique (SMOTE) diterapkan [17]. Sebelum SMOTE dijalankan, fitur Mel Spectrogram berbentuk 2D yang telah dihasilkan pada tahap sebelumnya diratakan (flattened) menjadi vektor satu dimensi. SMOTE bekerja dengan cara mensintesis sampel data baru pada kelas minoritas berdasarkan kedekatan jarak tetangga terdekat, bukan sekadar mereplikasi data yang sudah ada. Setelah jumlah sampel antar kelas seimbang, fitur-fitur tersebut dikembalikan (reshaped) ke dimensi semula sebelum diumpankan ke dalam model pembelajaran mendalam.

E. Arsitektur Model Hybrid CNN-LSTM

Arsitektur hibrida dirancang untuk memanfaatkan kekuatan dua metode deep learning yang berbeda secara sekuensial. Arsitektur lengkap dari model ini terdiri dari komponen-komponen berikut:

1. Lapisan CNN (Convolutional Neural Network): Bertindak sebagai ekstraktor fitur spasial lokal. Lapisan ini terdiri dari beberapa lapis Convolutional 2D yang dikombinasikan dengan Batch Normalization untuk mempercepat konvergensi, fungsi aktivasi ReLU, dan Max Pooling 2D untuk mereduksi dimensi spasial serta mengambil fitur tekstur spektral yang paling dominan dari gambar spectrogram.
2. Lapisan Reshape/TimeDistributed: Hasil fitur dari lapisan CNN kemudian diubah strukturnya agar memiliki dimensi urutan waktu (temporal timesteps), sehingga siap diproses oleh lapisan rekuren [18].
3. Lapisan LSTM (Long Short-Term Memory): Bertindak sebagai pembelajar sekuensial temporal. Lapisan LSTM menerima urutan fitur yang diekstrak oleh CNN untuk mempelajari bagaimana pola frekuensi tangisan bayi tersebut berevolusi atau berubah dari detik ke detik [19].
4. Lapisan Klasifikasi (Dense Layer): Fitur akhir dari LSTM dilewatkan melalui lapisan Dropout untuk mencegah overfitting, kemudian dihubungkan ke lapisan Dense akhir

dengan fungsi aktivasi Softmax untuk menghasilkan distribusi probabilitas prediksi terhadap lima kelas tangisan bayi [20].

F. Metrik Evaluasi Performa Model

Evaluasi ini didasarkan pada empat parameter kondisi prediksi, yaitu *True Positive* (TP) untuk data positif yang diprediksi benar, *True Negative* (TN) untuk data negatif yang diprediksi benar, *False Positive* (FP) untuk data negatif yang diprediksi sebagai positif, dan *False Negative* (FN) untuk data positif yang diprediksi sebagai negative [21]. Perhitungan matematis untuk masing-masing metrik evaluasi dijabarkan sebagai berikut:

1. Akurasi (Accuracy): Mengukur rasio total prediksi benar (baik positif maupun negatif) terhadap keseluruhan data yang diuji.

$$Akurasi = \frac{TP+TN}{TP+TN+FP+FN} \quad (3)$$

2. Presisi (Precision): Mengukur tingkat keakuratan antara data yang diminta dengan hasil prediksi yang dikeluarkan oleh model. Metrik ini menunjukkan kemampuan model dalam menghindari kesalahan klasifikasi positif yang keliru (false alarm).

$$Precision = \frac{TP}{TP+FP} \quad (4)$$

3. Recall: Mengukur keberhasilan model dalam menemukan kembali seluruh informasi atau sampel dari kelas target yang sebenarnya. Metrik ini sangat krusial untuk memastikan tidak ada jenis tangisan tertentu yang terlewat dari identifikasi model.

$$Recall = \frac{TP}{TP+FN} \quad (5)$$

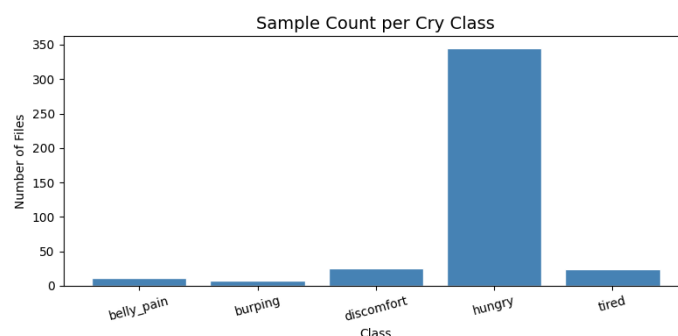
4. F1-Score: Merupakan nilai rata-rata harmonis (harmonic mean) yang menggabungkan metrik presisi dan recall. Metrik ini memberikan gambaran performa yang paling objektif dan seimbang, terutama ketika menghadapi karakteristik data yang memiliki potensi ketidakseimbangan kelas (imbalanced data).

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (6)$$

III. HASIL DAN PEMBAHASAN

A. Analisis Preprocessing Dataset dan Ekstraksi Fitur

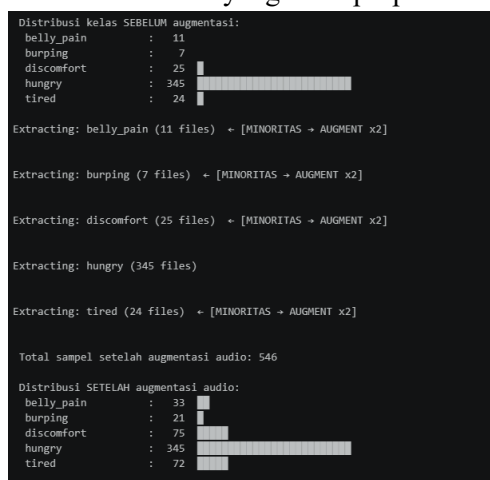
Hasil analisis menunjukkan bahwa distribusi data antar kelas tidak sepenuhnya seimbang. Beberapa kelas memiliki jumlah sampel yang lebih dominan dibandingkan kelas lainnya Seperti pada gambar 2.



Gambar 2. Dataset setiap Kelas

Berdasarkan analisis distribusi dataset, terdapat ketidakseimbangan yang signifikan dalam jumlah sampel antar kelas. Sebelum augmentasi, kelas lapar mendominasi dataset dengan 345 sampel (83,7% dari total data), sedangkan kelas sakit perut, bersendawa, ketidaknyamanan, dan lelah hanya memiliki 11, 7, 25, dan 24 sampel, masing-masing. Ketidakseimbangan distribusi ini berpotensi menyebabkan model lebih fokus pada pembelajaran pola di kelas mayoritas dan mengurangi kemampuannya untuk mengenali karakteristik kelas minoritas.

Untuk mengatasi masalah ini, hasil augmentasi meningkatkan jumlah titik data dari 412 menjadi 546 sampel, dengan jumlah sampel untuk kelas sakit perut meningkat menjadi 33, bersendawa menjadi 21, ketidaknyamanan menjadi 75, dan lelah menjadi 72, sedangkan kelas lapar tetap 345 sampel. Setelah penambahan data, dominasi kelas lapar menurun dari 83,7% menjadi 63,2%, menghasilkan distribusi data yang lebih proporsional.



Gambar 3. Ekstraksi dan Augmentasi Dataset

Kondisi pada Gambar 3 memberikan representasi karakteristik akustik yang lebih seimbang di setiap kategori tangisan bayi dan diharapkan dapat meningkatkan kemampuan generalisasi model CNN-LSTM dalam proses klasifikasi.

B. Penyeimbangan Dataset Menggunakan SMOTE

Ketidakseimbangan data (class imbalance) dapat menyebabkan model cenderung lebih baik dalam mengenali kelas mayoritas dibandingkan kelas minoritas, sehingga berpotensi menurunkan performa klasifikasi secara keseluruhan. Oleh karena itu, penelitian ini menerapkan metode Synthetic Minority Over-sampling Technique (SMOTE) untuk meningkatkan jumlah sampel pada kelas minoritas dan menghasilkan distribusi data yang lebih seimbang sebelum proses pelatihan model dilakukan.

Tabel 1. Perbandingan Distribusi Kelas Sebelum dan Sesudah Smote

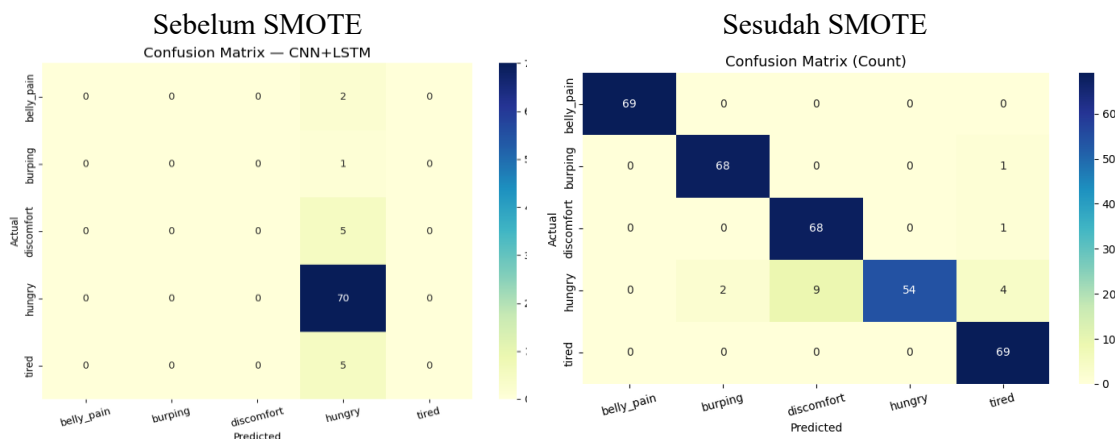
Sebelum SMOTE	Sesudah SMOTE
Shape sebelum SMOTE : X=(546, 216, 128, 1), y=(546,)	Shape setelah SMOTE : X=(1725, 216, 128, 1), y=(1725,)
Distribusi sebelum SMOTE: {np.int64(0): np.int64(33), np.int64(1): np.int64(21), np.int64(2): np.int64(75),	Distribusi setelah SMOTE: {np.int64(0): np.int64(345), np.int64(1): np.int64(345), np.int64(2): np.int64(345),

np.int64(3): np.int64(345), np.int64(4): np.int64(72)}	np.int64(3): np.int64(345), np.int64(4): np.int64(345)}
---	--

Tabel 1 memperlihatkan distribusi data sebelum dan sesudah penerapan SMOTE. Sebelum proses oversampling dilakukan, terdapat perbedaan jumlah sampel yang cukup signifikan antar kelas. Setelah SMOTE diterapkan, jumlah sampel pada setiap kelas menjadi lebih seimbang sehingga model memperoleh kesempatan yang sama untuk mempelajari karakteristik masing-masing kategori tangisan bayi. Keseimbangan distribusi data ini diharapkan mampu mengurangi bias model terhadap kelas mayoritas serta meningkatkan kemampuan klasifikasi pada kelas yang sebelumnya memiliki jumlah data terbatas.

C. Hasil Analisis Perbandingan Matrix Evaluasi.

Berdasarkan hasil evaluasi pada gambar 4, model yang dilatih menggunakan dataset hasil SMOTE menunjukkan peningkatan performa dibandingkan model yang dilatih menggunakan dataset asli. Kondisi ini terlihat dari nilai recall dan F1-score yang tidak merata antar kelas, serta adanya sejumlah kesalahan klasifikasi pada confusion matrix.



Gambar 4. Evaluasi Confusion Matrix sebelum dan sesudah SMOTE

Sebaliknya, setelah penerapan SMOTE, distribusi data antar kelas menjadi lebih seimbang sehingga model memperoleh representasi yang lebih baik terhadap karakteristik setiap kategori tangisan bayi. Hal ini tercermin pada peningkatan nilai precision, recall, dan F1-score pada beberapa kelas yang sebelumnya memiliki jumlah sampel terbatas. Selain itu, pola confusion matrix menunjukkan berkurangnya jumlah kesalahan prediksi pada kelas minoritas, yang menandakan bahwa model mampu mengenali karakteristik kelas tersebut dengan lebih baik.

Hasil tersebut menunjukkan bahwa penerapan SMOTE memberikan kontribusi positif terhadap proses pembelajaran model, terutama dalam mengatasi masalah ketidakseimbangan data yang umum ditemukan pada dataset tangisan bayi. Dengan distribusi data yang lebih proporsional, model dapat menghasilkan performa klasifikasi yang lebih stabil dan representatif pada seluruh kategori tangisan bayi.

D. Analisis Performa Kinerja Arsitektur

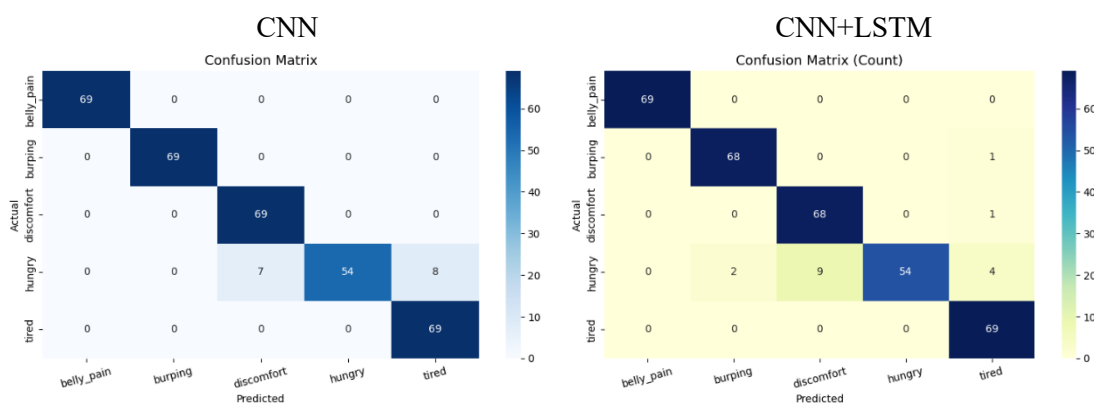
Untuk mengetahui pengaruh penanganan ketidakseimbangan data terhadap kinerja model, dilakukan perbandingan antara arsitektur CNN-LSTM yang dilatih menggunakan dataset asli dan dataset yang telah diseimbangkan menggunakan Synthetic Minority Over-sampling Technique (SMOTE). Hasil pengujian menunjukkan bahwa model CNN-LSTM tanpa SMOTE memperoleh nilai test accuracy sebesar 84,34% dengan test loss sebesar 0,6371. Setelah

penerapan SMOTE, akurasi model meningkat menjadi 95,07%. Berikut hasil pengujian pada Tabel 2

Tabel 2. Perbandingan Performa Model Sebelum dan Sesudah di SMOTE

HASIL	Arsitektur CNN+LSTM	
	Sebelum	Sesudah
Test Accuracy	84,34%	95,07%
Test Loss	0,6371	0,1160

Setelah penerapan SMOTE, jumlah sampel pada kelas minoritas bertambah melalui pembentukan data sintetis berdasarkan karakteristik data asli. Distribusi data yang lebih seimbang memungkinkan model memperoleh representasi yang lebih baik terhadap seluruh kategori tangisan bayi selama proses pelatihan. Kondisi ini berdampak pada peningkatan kemampuan generalisasi model yang ditunjukkan oleh kenaikan akurasi sebesar 10,73%.



Gambar 4. Confusion Matrix dan F1-Score Per Kelas Model CNN dan CNN+LSTM

Keberhasilan pencapaian akurasi di 95% pada kedua model berbasis konvolusi ini juga tidak lepas dari kontribusi tahapan penanganan data. Pada Gambar 4 penerapan teknik augmentasi audio (seperti pemberian noise dan pitch shifting) berpadu dengan metode SMOTE terbukti mampu menghilangkan bias kelas mayoritas. Penyeimbangan jumlah sampel pada lima kategori tangisan memastikan bahwa model mempelajari karakteristik unik dari kelas-kelas minoritas seperti 'bersendawa' atau 'lelah' dengan porsi yang sama baiknya dengan kelas mayoritas seperti 'lapar'.

IV. KESIMPULAN

Berdasarkan hasil pengujian dan analisis yang telah dilakukan, dapat disimpulkan bahwa:

1. Implementasi arsitektur hibrida CNN-LSTM dengan ekstraksi fitur Mel Spectrogram terbukti efektif dan akurat dalam mengklasifikasikan lima jenis suara tangisan bayi.
2. Penerapan Synthetic Minority Over-sampling Technique (SMOTE) berhasil meningkatkan kualitas data pelatihan dengan menghasilkan distribusi kelas yang lebih seimbang. Hasil analisis menunjukkan bahwa model CNN-LSTM yang dilatih menggunakan data SMOTE mencapai kinerja yang lebih baik daripada sebelum penerapan SMOTE, seperti yang ditunjukkan oleh peningkatan akurasi klasifikasi dan nilai presisi, recall, dan F1-score yang lebih konsisten di semua kategori tangisan bayi..

3. Penelitian ini berkontribusi dalam pengembangan arsitektur deep learning yang tangguh terhadap variasi durasi audio, sebagai landasan untuk mewujudkan sistem asisten pengasuhan bayi cerdas di masa depan.

UCAPAN TERIMAKASIH

Penulis mengucapkan terima kasih kepada Program Studi Teknik Informatika Universitas Nusantara PGRI Kediri yang telah memberikan dukungan akademik selama pelaksanaan penelitian ini. Penulis juga menyampaikan apresiasi kepada dosen pembimbing dan seluruh pihak yang telah memberikan masukan, arahan, serta dukungan dalam proses penyusunan penelitian hingga artikel ini dapat diselesaikan dengan baik.

Selain itu, penulis mengucapkan terima kasih kepada pengembang dataset *Donate-a-Cry Corpus* yang telah menyediakan data penelitian secara terbuka sehingga penelitian mengenai klasifikasi tangisan bayi ini dapat dilaksanakan.

DAFTAR PUSTAKA

- [1] L. Yunita and D. Surayana, "Perkembangan personality sosial usia bayi dan toddler," *Jurnal Family Education*, vol. 1, no. 4, pp. 14–22, 2021. <https://doi.org/10.24036/jfe.v1i4.20>
- [2] H. Yoo, E. H. Buder, D. D. Bowman, G. M. Bidelman, and D. K. Oller, "Acoustic correlates and adult perceptions of distress in infant speech-like vocalizations and cries," *Front. Psychol.*, vol. 10, p. 1154, 2019. <https://doi.org/10.3389/fpsyg.2019.01154>
- [3] D. E. Schoth and C. Lioffi, "A systematic review of experimental paradigms for exploring biased interpretation of ambiguous information with emotional and neutral associations," *Front. Psychol.*, vol. 8, p. 171, 2017. <https://doi.org/10.1371/journal.pone.0318296>
- [4] C. Li, A. Pourtaherian, L. Van Onzenoort, W. E. T. a Ten, and P. H. N. de With, "Infant monitoring system for real-time and remote discomfort detection," *IEEE Transactions on Consumer Electronics*, vol. 66, no. 4, pp. 336–345, 2020. DOI: 10.1109/TCE.2020.3031359
- [5] A. S. H. George and A. S. George, "Decoding Infant Communication: Understanding the Meaning Behind Baby's Cries," *Partners Universal Innovative Research Publication*, vol. 1, no. 1, pp. 1–14, 2023. <https://doi.org/10.5281/zenodo.8432700>
- [6] T. H. Rochadiani, "Pendekatan Transfer Learning Untuk Klasifikasi Tangisan Bayi Dengan Imbalance Dataset," *The Indonesian Journal of Computer Science*, vol. 13, no. 2, 2024. <https://doi.org/10.33022/ijcs.v13i2.3834>
- [7] P. Cariani and J. M. Baker, "Time is of the essence: neural codes, synchronies, oscillations, architectures," *Front. Comput. Neurosci.*, vol. 16, p. 898829, 2022. <https://doi.org/10.3389/fncom.2022.898829>
- [8] M. Tiwari and D. K. Verma, "Enhanced text-independent speaker recognition using MFCC, Bi-LSTM, and CNN-based noise removal techniques," *Int. J. Speech Technol.*, vol. 27, no. 4, pp. 1013–1026, 2024. <https://doi.org/10.1007/s10772-024-10150-4>
- [9] C. Asuai, A. P. Arinomor, C. T. Atumah, I. F. Kowhoro, and D. E. Ogheneochuko, "Hybrid CNN-LSTM Architectures for Deepfake Audio Detection Using Mel Frequency Cepstral Coefficients and Spectrogram Analysis," *American Journal of Mathematical and Computer Modelling*, vol. 10, no. 3, pp. 98–109, 2025. DOI: 10.11648/j.ajmcm.20251003.12

- [10] B. A. Febryanto and I. Tahyudin, "Perbandingan Algoritma CNN, LSTM, FNN untuk Diagnosa Fibrosis Hati dengan Citra Medis.," *Techno. com*, vol. 24, no. 1, 2025. Doi: 10.62411/tc.v24i1.12020
- [11] G. Meriyem, D. Abdelaziz, and E. Mohamed, "From Traditional to Deep Learning Methods for Baby Cry Audio Signal Classification: A Systematic Literature Review," in *2025 International Conference on Circuit, Systems and Communication (ICCSC)*, IEEE, 2025, pp. 1–6. DOI: 10.1109/ICCSC66714.2025.11135198
- [12] I. P. A. A. Mahendra and K. Kusriani, "Analisis Komparatif Klasifikasi Sentimen Pengguna Aplikasi Investasi Menggunakan Algoritma Hybrid Cnn-Lstm, Cnn-Gru Dengan Implementasi Smote," *Teknimedia: Teknologi Informasi dan Multimedia*, vol. 6, no. 1, pp. 112–118, 2025. <https://doi.org/10.46764/teknimedia.v6i1.258>
- [13] N. M. Susetio, "Analysis of the Effect of Data Augmentation on the Classification of Indonesian Pronunciation Accuracy in Pronouncing the English Word 'City' Using Deep Learning Methods with MEL Spectrogram and Convolutional Neural Network (CNN)," *Available at SSRN 5196866*, 2025. <http://dx.doi.org/10.2139/ssrn.5196866>
- [14] Haoran Kui, Jiahua Pan, Rong Zong, Hongbo Yang, Weilian Wang, "Heart sound classification based on log Mel-frequency spectral coefficients features and convolutional neural networks," *Biomedical Signal Processing and Control*, Volume 69, 2021, 102893, ISSN 1746-8094, <https://doi.org/10.1016/j.bspc.2021.102893>
- [15] A. Ustubioglu, B. Ustubioglu, and G. Ulutas, "Mel spectrogram-based audio forgery detection using CNN," *Signal Image Video Process.*, vol. 17, no. 5, pp. 2211–2219, 2023. <https://doi.org/10.1007/s11760-022-02436-4>
- [16] C. Aliferis and G. Simon, "Overfitting, underfitting and general model overconfidence and under-performance pitfalls and best practices in machine learning and AI," *Artificial intelligence and machine learning in health care and medical sciences: Best practices and pitfalls*, pp. 477–524, 2024. https://doi.org/10.1007/978-3-031-39355-6_10
- [17] D. Fadillah, E. Haerani, F. Wulandari, and F. Syafria, "Klasifikasi Kondisi Janin Menggunakan Algoritma K-Nearest Neighbors dan Teknik SMOTE Berdasarkan Data Kardiotogram," *Bulletin of Computer Science Research*, vol. 5, no. 4, pp. 482–489, 2025. <https://doi.org/10.47065/bulletinscr.v5i4.585>
- [18] T. Ahmad, J. Wu, H. S. Alwageed, F. Khan, J. Khan, and Y. Lee, "Human activity recognition based on deep-temporal learning using convolution neural networks features and bidirectional gated recurrent unit with features selection," *IEEE access*, vol. 11, pp. 33148–33159, 2023. doi: 10.1109/ACCESS.2023.3263155
- [19] Lusi Dwi Anggraini, Umi Mahdiyah, Resty Wulanningrum, "Pengujian Simple CNN Dengan Arsitektur LeNet5 Pada Penyakit Daun Jagung", *inotek*, vol. 9, no. 1, pp. 711–719, Jul. 2025, doi: 10.29407/9qdmem64.
- [20] A. Farzad, H. Mashayekhi, and H. Hassanpour, "A comparative performance analysis of different activation functions in LSTM networks for classification," *Neural Comput. Appl.*, vol. 31, no. 7, pp. 2507–2521, 2019. <https://doi.org/10.1007/s00521-017-3210-6>
- [21] M. C. Rani, R. A. Dewi, F. D. Azkia, M. Wahyudi, and A. S. Budiman, "Perbandingan Algoritma Random Forest, Naive Bayes, Dan Neural Network Dalam Klasifikasi Penyakit Jantung," *Jurnal Sains Informatika Terapan*, vol. 4, no. 2, pp. 77–84, 2025. <https://doi.org/10.62357/jsit.v4i2.609>