ANALISIS KOMPARATIF KINERJA MODEL *NEURAL NETWORK* RINGAN (CNN, LSTM, *TRANSFORMER*) UNTUK DETEKSI *DEEPFAKE AUDIO* PADA *DATASET* WAVEFAKE

SKRIPSI

diajukan untuk memenuhi sebagian syarat untuk memperoleh gelar Sarjana Teknik pada Program Studi Mekatronika dan Kecerdasan Buatan

Oleh:
Rafharum Fatimah
NIM 2104428

PROGRAM STUDI MEKATRONIKA DAN KECERDASAN BUATAN
KAMPUS UPI DI PURWAKARTA
UNIVERSITAS PENDIDIKAN INDONESIA
TAHUN 2025

LEMBAR HAK CIPTA

ANALISIS KOMPARATIF KINERJA MODEL *NEURAL NETWORK* RINGAN (CNN, LSTM, *TRANSFORMER*) UNTUK DETEKSI *DEEPFAKE AUDIO* PADA *DATASET* WAVEFAKE

Oleh
Rafharum Fatimah

Sebuah skripsi yang diajukan untuk memenuhi salah satu syarat memperoleh gelar Sarjana Teknik pada Program Studi Mekatronika dan Kecerdasan Buatan

Universitas Pendidikan Indonesia
Agustus 2025

LEMBAR PENGESAHAN

Rafharum Fatimah

2104428

Analisis Komparatif Kinerja Model *Neural Network* Ringan (CNN, LSTM, *Transformer*) untuk Deteksi *Deepfake Audio* pada *Dataset* WaveFake

Disetujui dan disahkan oleh pembimbing,

Pembimbing I,

Mahmudah Salwa Gianti, S.Si., M.Eng.

NIP. 920210919960408201

Pembimbing II,

Muhammad Rizalul Wahid, S.Si., M.T.

NIP. 920210919940401101

Mengetahui

Ketua Program Studi

Mekatronika dan Kecerdasan Buatan

Dewi Indriati Hadi Putri, S.Pd., M.T.

NIP. 920190219900126201

ABSTRACT

The emergence of deepfake audio as a result of text-to-speech-based voice manipulation poses a serious threat to digital security. Therefore, this study aims to compare the performance of three lightweight deep learning models, namely Convolutional Neural Network (CNN), Long Short-Term Memory (LSTM), and Transformer, in detecting deepfake audio using the WaveFake dataset. The dataset was processed through a preprocessing stage and extracted into three main features: MFCC, Mel-spectrogram, and spectrogram. These were then used as input for each model with uniform training parameters. Evaluation results showed that CNN achieved the highest accuracy at 92.8%, followed by LSTM at 87.8%, while Transformer obtained 84.9%. CNN excels due to its ability to extract local patterns in audio data, while Transformer still require further optimization. LSTM is relatively optimal in handling complex spectral dimensions but not as effective as CNN. This study concludes that CNN is the most effective architecture for detecting deepfake audio on the WaveFake dataset and has the potential to be applied in efficient digital security systems.

Keywords: CNN, *Deepfake*, LSTM, SVM, *Transformer*.

DAFTAR ISI

DAFTAR TABEL

DAFTAR GAMBAR

DAFTAR LAMPIRAN

DAFTAR PUSTAKA

Ajinurseto, G., Bakrim, L. O., & Islamuddin, N. (2023). Penerapan Metode Mel Frequency Cepstral Coefficients pada Sistem Pengenalan Suara Berbasis Desktop. *Infomatek*, *25*(1), 11–20. https://doi.org/10.23969/infomatek.v25i1.6109

Ali, G., Rashid, J., Rameez Ul Hussnain, M., Usman Tariq, M., Ghani, A., & Kwak, D. (2024). Beyond the Illusion: Ensemble Learning for Effective Voice Deepfake Detection. *IEEE Access*, *12*, 149940–149959. https://doi.org/10.1109/ACCESS.2024.3457866

Allamy, S., & Koerich, A. L. (2021). *1D CNN Architectures for Music Genre Classification* (arXiv:2105.07302). arXiv. https://doi.org/10.48550/arXiv.2105.07302

Andayani, F., Theng, L. B., Tsun, M. T., & Chua, C. (2022). Hybrid LSTM-Transformer Model for Emotion Recognition From Speech Audio Files. *IEEE Access*, *10*, 36018–36027. https://doi.org/10.1109/ACCESS.2022.3163856

Anggraini, D., Gamayanto, I., & Wibowo, S. (2025). Comparing Decision Tree and Support Vector Machines in Hospital Satisfaction. *Journal of Applied Informatics and Computing*, *9*(2), 364–372. https://doi.org/10.30871/jaic.v9i2.9203

Aravind Chinnaraju. (2025). Benchmarking cross-platform AI: Web Assembly, ONNX Runtime and TVM for Real-Time Web, Mobile, and IoT Deployment. *World Journal of Advanced Research and Reviews*, *26*(2), 1937–1963. https://doi.org/10.30574/wjarr.2025.26.2.1832

Ashar, A., Bhatti, M. S., & Mushtaq, U. (2020). Speaker Identification Using a Hybrid CNN-MFCC Approach. *2020 International Conference on Emerging Trends in Smart Technologies (ICETST)*, 1–4. https://doi.org/10.1109/ICETST49965.2020.9080730

Ashfaq, S., AskariHemmat, M., Sah, S., Saboori, E., Mastropietro, O., & Hoffman, A. (n.d.). *Accelerating Deep Learning Model Inference on Arm CPUs with Ultra-Low Bit Quantization and Runtime*.

Bajwa, M. K. Z., Castiglione, A., & Pero, C. (2025). Mel Spectrogram-Based CNN Framework for Explainable Audio Deepfake Detection. In L. Barolli (Ed.), *Advanced Information Networking and Applications* (Vol. 252, pp. 407–416). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-87784-1_37

Besta, M., & Hoefler, T. (2023). *Parallel and Distributed Graph Neural Networks: An In-Depth Concurrency Analysis* (arXiv:2205.09702). arXiv. https://doi.org/10.48550/arXiv.2205.09702

Beyer, L., Zhai, X., & Kolesnikov, A. (2022). *Better plain ViT baselines for ImageNet-1k* (arXiv:2205.01580). arXiv. https://doi.org/10.48550/arXiv.2205.01580

Bisogni, C., Loia, V., Nappi, M., & Pero, C. (2024). Acoustic features analysis for explainable machine learning-based audio spoofing detection. *Computer Vision and Image Understanding*, *249*, 104145. https://doi.org/10.1016/j.cviu.2024.104145

Carson, A., Välimäki, V., Wright, A., & Bilbao, S. (2025). *Resampling Filter Design for Multirate Neural Audio Effect Processing* (arXiv:2501.18470). arXiv. https://doi.org/10.48550/arXiv.2501.18470

Chang, S., Park, H., Cho, J., Park, H., Yun, S., & Hwang, K. (2021). *SubSpectral Normalization for Neural Audio Data Processing* (arXiv:2103.13620). arXiv. https://doi.org/10.48550/arXiv.2103.13620

Chao, L., Tao, J., Yang, M., Li, Y., & Wen, Z. (n.d.). *Audio Visual Emotion Recognition with Temporal Alignment and Perception Attention*.

Charanteja, B. G., David, A., Sasank, B., Abhishek, B., & Sudhakara Rao, A. V. S. (2023). Noise removal from audio using CNN. International Journal of Recent Scientific Research, 14(11), 4306–4311. https://doi.org/10.24327/ijrsr.20231411.0807

Chaturvedi, K., Gasthuber, J., & Abdelaal, M. (2025a). *EdgeMLOps: Operationalizing ML models with Cumulocity IoT and thin-edge.io for Visual quality Inspection*.

Chaturvedi, K., Gasthuber, J., & Abdelaal, M. (2025b). *EdgeMLOps: Operationalizing ML models with Cumulocity IoT and thin-edge.io for Visual quality Inspection* (arXiv:2501.17062). arXiv. https://doi.org/10.48550/arXiv.2501.17062

Chen, H., & Chang, X. (2021). Photovoltaic power prediction of LSTM model based on Pearson feature selection. *Energy Reports*, *7*, 1047–1054. https://doi.org/10.1016/j.egyr.2021.09.167

Chernyavskiy, A., Ilvovsky, D., & Nakov, P. (2021). *Transformers: "The End of History" for NLP?* (arXiv:2105.00813). arXiv. https://doi.org/10.48550/arXiv.2105.00813

Dastres, R., & Soori, M. (n.d.). *Artificial Neural Network Systems*.

Dhindsa, A., Bhatia, S., Agrawal, S., & Sohi, B. S. (2021). An Improvised Machine Learning Model Based on Mutual Information Feature Selection Approach for Microbes Classification. *Entropy*, *23*(2), 257. https://doi.org/10.3390/e23020257

Drewek-Ossowicka, A., Pietrołaj, M., & Rumiński, J. (2021). A survey of neural networks usage for intrusion detection systems. *Journal of Ambient Intelligence and Humanized Computing*, *12*(1), 497–514. https://doi.org/10.1007/s12652-020-02014-x

Edwards, P., Nebel, J.-C., Greenhill, D., & Liang, X. (2024). A Review of Deepfake Techniques: Architecture, Detection, and Datasets. *IEEE Access*, *12*, 154718–154742. https://doi.org/10.1109/ACCESS.2024.3477257

Feng, X., Wang, Y., & Luo, F. (2024). Application of AI-powered audio noise reduction system in remote education. In Proceedings of the 2024 2nd International Conference on Information Education and Artificial Intelligence (ICIEAI 2024) (pp. 1–5). ACM. https://doi.org/10.1145/3724504.3724610

Frank, J., & Schönherr, L. (2021). WaveFake: A Data Set to Facilitate Audio Deepfake Detection. *Advances in Neural Information Processing Systems*, *NeurIPS*.

Géron, A. (2019). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems* (2nd ed.). O'Reilly Media.

Gong, Y., Lai, C. I. J., Chung, Y. A., & Glass, J. (2022a). SSAST: Self-Supervised Audio Spectrogram Transformer. *Proceedings of the 36th AAAI Conference on Artificial Intelligence, AAAI 2022*, *36*, 10699–10709. https://doi.org/10.1609/aaai.v36i10.21315

Gong, Y., Lai, C.-I. J., Chung, Y.-A., & Glass, J. (2022b). *SSAST: Self-Supervised Audio Spectrogram Transformer* (arXiv:2110.09784). arXiv. https://doi.org/10.48550/arXiv.2110.09784

Groumpos, P. P. (2023). A Critical Historic Overview of Artificial Intelligence: Issues, Challenges, Opportunities, and Threats. *Artificial Intelligence and Applications*, *1*(4), 181–197. https://doi.org/10.47852/bonviewAIA3202689

Gupta, J., Pathak, S., & Kumar, G. (2022). Deep Learning (CNN) and Transfer Learning: A Review. *Journal of Physics: Conference Series*, *2273*(1), 012029. https://doi.org/10.1088/1742-6596/2273/1/012029

Hamza, A., Javed, A. R. R., Iqbal, F., Kryvinska, N., Almadhor, A. S., Jalil, Z., & Borghol, R. (2022). Deepfake Audio Detection via MFCC Features Using Machine Learning. *IEEE Access*, *10*, 134018–134028. https://doi.org/10.1109/ACCESS.2022.3231480

Hashmi, A., Shahzad, S. A., Lin, C.-W., Tsao, Y., & Wang, H.-M. (2024). *Unmasking Illusions: Understanding Human Perception of Audiovisual Deepfakes* (arXiv:2405.04097). arXiv. https://doi.org/10.48550/arXiv.2405.04097

Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., & Adam, H. (2017). *MobileNets: Efficient Convolutional*

Neural Networks for Mobile Vision Applications (arXiv:1704.04861). arXiv. https://doi.org/10.48550/arXiv.1704.04861

Hua, T. K. (2022). *A Short Review on Machine Learning*. https://doi.org/10.22541/au.166490976.66390273/v1

Iqbal, F., Abbasi, A., Javed, A. R., Jalil, Z., & Al-Karaki, J. (n.d.). *Deepfake Audio Detection via Feature Engineering and Machine Learning*.

Jajal, P., Jiang, W., Tewari, A., Kocinare, E., Woo, J., Sarraf, A., Lu, Y.-H., Thiruvathukal, G. K., & Davis, J. C. (2024). Interoperability in Deep Learning: A User Survey and Failure Analysis of ONNX Model Converters. *Proceedings of the 33rd ACM SIGSOFT International Symposium on Software Testing and Analysis*, 1466–1478. https://doi.org/10.1145/3650212.3680374

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning* (Vol. 103). Springer New York. https://doi.org/10.1007/978-1-4614-7138-7

Jiang, M. (2025). Support Vector Machines and Their Application in Mechanical Fault Diagnosis. *Theoretical and Natural Science*, *95*(1), 136–145. https://doi.org/10.54254/2753-8818/2024.21346

Karnouskos, S. (2020). Artificial Intelligence in Digital Media: The Era of Deepfakes. *IEEE Transactions on Technology and Society*, *1*(3), 138–147. https://doi.org/10.1109/TTS.2020.3001312

Kobrinskii, B. A. (2023). Artificial Intelligence: Problems, Solutions, and Prospects. *Pattern Recognition and Image Analysis*, *33*(3), 217–220. https://doi.org/10.1134/S1054661823030203

Kondaveeti, H. K., Vatsavayi, V. K., Mangapathi, S., & Yasaswini, R. M. (2023). Lightweight Deep Learning: Introduction, Advancements, and Applications. In R. Gharoie Ahangar & M. Napier (Eds.), *Advances in Business Information Systems and Analytics* (pp. 242–259). IGI Global. https://doi.org/10.4018/978-1-6684-8386-2.ch012

Le, T. D. N., Teh, K. K., & Tran, H. D. (2024). *Continuous Learning of Transformer-based Audio Deepfake Detection* (arXiv:2409.05924). arXiv. https://doi.org/10.48550/arXiv.2409.05924

Li, S., Deng, Y. Q., Zhu, Z. L., Hua, H. L., & Tao, Z. Z. (2021). A comprehensive review on radiomics and deep learning for nasopharyngeal carcinoma imaging. In *Diagnostics* (Vol. 11, Issue 9). https://doi.org/10.3390/diagnostics11091523

Liu, B., Liu, B., Zhu, T., & Ding, M. (2025). A Review of Deepfake and Its Detection: From Generative Adversarial Networks to Diffusion Models. *International Journal of Intelligent Systems*, *2025*(1), 9987535. https://doi.org/10.1155/int/9987535

Liu, W., She, T., Liu, J., Li, B., Yao, D., Liang, Z., & Wang, R. (n.d.). *Lips Are Lying: Spotting the Temporal Inconsistency between Audio and Visual in Lip-Syncing DeepFakes*.

Long, Y. (2025). Enhanced SVM-based model for predicting cyberspace vulnerabilities: Analyzing the role of user group dynamics and capital influx. *Plos One*, *20*(7 July), 1–20. https://doi.org/10.1371/journal.pone.0327476

Madhiarasan, M., & Louzazni, M. (2022). Analysis of Artificial Neural Network: Architecture, Types, and Forecasting Applications. *Journal of Electrical and Computer Engineering*, *2022*, 1–23. https://doi.org/10.1155/2022/5416722

Mahmud, B. U., & Sharmin, A. (n.d.). *Deep Insights of Deepfake Technology: A Review*.

Mamun, N., Zilany, M. S. A., Hansen, J. H. L., & Davies-Venn, E. E. (2021). An intrusive method for estimating speech intelligibility from noisy and distorted signals. The Journal of the Acoustical Society of America, 150(3), 1762–1778. https://doi.org/10.1121/10.0005899

Masood, M., Nawaz, M., Malik, K. M., Javed, A., Irtaza, A., & Malik, H. (2023). Deepfakes generation and detection: State-of-the-art, open challenges,

countermeasures, and way forward. *Applied Intelligence*, *53*(4), 3974–4026. https://doi.org/10.1007/s10489-022-03766-z

Mubarak, R., Alsboui, T., Alshaikh, O., Inuwa-Dutse, I., Khan, S., & Parkinson, S. (2023). A Survey on the Detection and Impacts of Deepfakes in Visual, Audio, and Textual Formats. *IEEE Access*, *11*, 144497–144529. https://doi.org/10.1109/ACCESS.2023.3344653

Muruganandham, P., Thangasamy, G. R., Jayaraman, S., & Dharmarajan, R. (2025). LSTM autoencoder based parallel architecture for deepfake audio detection with dynamic residual encoding and feature fusion. *Scientific Reports*, *15*(1), 23514. https://doi.org/10.1038/s41598-025-08198-6

Panjaitan, F., Simbolon, R. S., & Batubara, J. (2025). Handwritten Digit Recognition Using Support Vector Machine (SVM) with Radial Basis Function (RBF) Kernel. *Journal Basic Science and Technology*, *14*(1), 10–18.

Patel, Y., Tanwar, S., Gupta, R., Bhattacharya, P., Davidson, I. E., Nyameko, R., Aluvala, S., & Vimal, V. (2023). Deepfake Generation and Detection: Case Study and Challenges. *IEEE Access*, *11*(October), 143296–143323. https://doi.org/10.1109/ACCESS.2023.3342107

Pham, L., Lam, P., Nguyen, T., Nguyen, H., & Schindler, A. (2024). *Deepfake Audio Detection Using Spectrogram-based Feature and Ensemble of Deep Learning Models* (arXiv:2407.01777). arXiv. https://doi.org/10.48550/arXiv.2407.01777

Pons, J., Pascual, S., Cengarle, G., & Serrà, J. (2021). *Upsampling artifacts in neural audio synthesis* (arXiv:2010.14356). arXiv. https://doi.org/10.48550/arXiv.2010.14356

Pudasaini, A., Al-Hawawreh, M., Bouadjenek, M. R., Hacid, H., & Aryal, S. (2025). A comprehensive study of audio profiling: Methods, applications, challenges, and future directions. Neurocomputing, 640, 130334. https://doi.org/10.1016/j.neucom.2025.130334

Purwono, P., Ma'arif, A., Rahmaniar, W., Fathurrahman, H. I. K., Frisky, A. Z. K., & Haq, Q. M. U. (2023). Understanding of Convolutional Neural Network

(CNN): A Review. *International Journal of Robotics and Control Systems*, *2*(4), 739–748. https://doi.org/10.31763/ijrcs.v2i4.888

Rainio, O., Teuho, J., & Klén, R. (2024). Evaluation metrics and statistical tests for machine learning. *Scientific Reports*, *14*(1), 6086. https://doi.org/10.1038/s41598-024-56706-x

Razavi, S. (2021). Deep learning, explained: Fundamentals, explainability, and bridgeability to process-based modelling. *Environmental Modelling & Software*, *144*, 105159. https://doi.org/10.1016/j.envsoft.2021.105159

Salvi, D., Hosler, B., Bestagini, P., Stamm, M. C., & Tubaro, S. (2023). TIMIT-TTS: A Text-to-Speech Dataset for Multimodal Synthetic Media Detection. *IEEE Access*, *11*, 50851–50866. https://doi.org/10.1109/ACCESS.2023.3276480

Schäfer, K., Choi, J. E., & Zmudzinski, S. (2024). Explore the World of Audio Deepfakes: A Guide to Detection Techniques for Non-Experts. *ACM International Conference Proceeding Series*, 13–22. https://doi.org/10.1145/3643491.3660289

Schoneveld, L., Othmani, A., & Abdelkawy, H. (2021). Leveraging Recent Advances in Deep Learning for Audio-Visual Emotion Recognition. *Pattern Recognition Letters*, *146*, 1–7. https://doi.org/10.1016/j.patrec.2021.03.007

Schwartz, R., Dodge, J., Smith, N. A., & Etzioni, O. (2020). Green AI. *Communications of the ACM, 63*(12), 54–63. https://doi.org/10.1145/3381831

Selvaraj, P., Maidin, S. S., & Yang, Q. (2025). Speech enhancement using sliding window empirical mode decomposition with median filtering technique. Journal of Applied Data Sciences, 6(1), 143–154. https://doi.org/10.47738/jads.v6i1.470

Shaaban, O. A., & Yildirim, R. (2025). Audio Deepfake Detection Using Deep Learning. *Engineering Reports*, *7*(3), e70087. https://doi.org/10.1002/eng2.70087

Shaaban, O. A., Yildirim, R., & Alguttar, A. A. (2023). Audio Deepfake Approaches. *IEEE Access*, *11*, 132652–132682. https://doi.org/10.1109/ACCESS.2023.3333866

Song, D., Lee, N., Kim, J., & Choi, E. (2024). Anomaly Detection of Deepfake Audio Based on Real Audio Using Generative Adversarial Network Model. *IEEE Access*, *12*(November), 184311–184326. https://doi.org/10.1109/ACCESS.2024.3506973

Sonobe, R., Takamichi, S., & Saruwatari, H. (2017). *JSUT corpus: free large-scale Japanese speech corpus for end-to-end speech synthesis*. 1–4. http://arxiv.org/abs/1711.00354

Taye, M. M. (2023). Theoretical Understanding of Convolutional Neural Network: Concepts, Architectures, Applications, Future Directions. *Computation*, *11*(3), 52. https://doi.org/10.3390/computation11030052

Todisco, M., Wang, X., Vestman, V., Sahidullah, Md., Delgado, H., Nautsch, A., Yamagishi, J., Evans, N., Kinnunen, T. H., & Lee, K. A. (2019). ASVspoof 2019: Future Horizons in Spoofed and Fake Audio Detection. *Interspeech 2019*, 1008–1012. https://doi.org/10.21437/Interspeech.2019-2249

Usama Tanveer Gujjar, M., Munir, K., Amjad, M., Rehman, A. U., & Bermak, A. (2024). Unmasking the Fake: Machine Learning Approach for Deepfake Voice Detection. *IEEE Access*, *12*, 197442–197453. https://doi.org/10.1109/ACCESS.2024.3521026

Vujovic, Ž. Đ. (2021). Classification Model Evaluation Metrics. *International Journal of Advanced Computer Science and Applications*, *12*(6). https://doi.org/10.14569/IJACSA.2021.0120670

Wright, L. G., Onodera, T., Stein, M. M., Wang, T., Schachter, D. T., Hu, Z., & McMahon, P. L. (2022). Deep physical neural networks trained with backpropagation. *Nature*, *601*(7894), 549–555. https://doi.org/10.1038/s41586-021-04223-6

Yang, S., Liu, A. T., & Lee, H. (2020). *Understanding Self-Attention of Self-Supervised Audio Transformers* (arXiv:2006.03265). arXiv. https://doi.org/10.48550/arXiv.2006.03265

Yang, S. W., Liu, A. T., & Lee, H. Y. (2020). Understanding self-attention of self-supervised audio transformers. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, *2020-October*, 3785–3789. https://doi.org/10.21437/Interspeech.2020-2231

Yi, J., Wang, C., Tao, J., Zhang, X., Zhang, C. Y., & Zhao, Y. (2023a). *Audio Deepfake Detection: A Survey*. *14*(8), 1–20.

Yi, J., Wang, C., Tao, J., Zhang, X., Zhang, C. Y., & Zhao, Y. (2023b). *Audio Deepfake Detection: A Survey* (arXiv:2308.14970). arXiv. https://doi.org/10.48550/arXiv.2308.14970

Yoneyama, R., Yamamoto, R., & Tachibana, K. (2023). *Nonparallel High-Quality Audio Super Resolution with Domain Adaptation and Resampling CycleGANs* (arXiv:2210.15887). arXiv. https://doi.org/10.48550/arXiv.2210.15887

Zaman, K., Sah, M., Direkoglu, C., & Unoki, M. (2023). A Survey of Audio Classification Using Deep Learning. *IEEE Access*, *11*, 106620–106649. https://doi.org/10.1109/ACCESS.2023.3318015

Zaman, K., Samiul, I. J. A. M., Sah, M., Direkoglu, C., Okada, S., & Unoki, M. (2024). Hybrid Transformer Architectures With Diverse Audio Features for Deepfake Speech Classification. *IEEE Access*, *12*, 149221–149237. https://doi.org/10.1109/ACCESS.2024.3478731

Zhou, H., Wang, X., & Zhu, R. (2022). Feature selection based on mutual information with correlation coefficient. *Applied Intelligence*, *52*(5), 5457–5474. https://doi.org/10.1007/s10489-021-02524-x