PROPOSAL TUGAS AKHIR

Evaluasi Komparatif Model Self-Supervised Wav2Vec2 dan Data2Vec untuk Retrieval Semantik Ayat Al-Qur'an Berbasis Audio

Disusun Oleh:

Mujahid Ansori Majid 1197050093

PROGRAM STUDI TEKNIK INFORMATIKA

FAKULTAS SAINS DAN TEKNOLOGI

UNIVERSITAS ISLAM NEGERI SUNAN GUNUNG DJATI

BANDUNG

2025

DAFTAR ISI

DAFTAR GAMBAR

DAFTAR TABEL

PENDAHULUAN

1. Latar Belakang

Perkembangan teknologi modern telah memengaruhi metode tradisional dalam penghafalan Al-Qur’an, mendorong munculnya pendekatan berbasis digital yang mengintegrasikan kemajuan dalam audio, interaktivitas, dan kecerdasan buatan.. Meski demikian, era digitalisasi juga membawa tantangan baru berupa elemen pengalih perhatian yang dapat mengganggu konsentrasi dalam proses menghafal [1]. Pergeseran pola kehidupan kontemporer turut memperparah situasi serta keterbatasan waktu yang dialokasikan untuk kegiatan spiritual, khususnya dalam proses penghafalan Al-Qur'an [2].

Riset ini didesain untuk menjadi salah satu solusi bagi permasalahan dalam hal tersebut, Penerapan metode rekomendasi ayat dalam proses penghafalan dapat menjadi alat bantu yang efektif sekaligus membentuk pendekatan holistik untuk mendukung pembelajar Al-Qur’an dalam mengatasi berbagai hambatan yang dihadapi. Sebagaimana dijelaskan dalam penelitian yang ditulis oleh Shaklawoon et al [3], penerapan teknologi inovatif seperti pengenalan suara (*speech recognition*) terbukti mampu membantu meningkatkan hasil penghafalan Al-Qur'an. Sistem yang mereka kembangkan berhasil mendeteksi kesalahan dalam urutan ayat dan pengucapan dengan tingkat akurasi yang tinggi, sehingga mendukung proses hafalan secara mandiri dan efektif [4].

Dengan merujuk pada potensi yang ditunjukkan dalam penelitian sebelumnya, studi ini diarahkan untuk mengeksplorasi dan mengevaluasi penerapan dua model *speech recognition* modern guna menghasilkan sistem yang lebih adaptif dalam mendukung proses hafalan Al-Qur’an. Penelitian ini mengimplementasikan model *Wav2vec2* dan *Data2Vec* sebagai instrumen analisis. Sasaran utama penelitian mencakup penyusunan rekomendasi implementasi teknologi yang produktif untuk menunjang aktivitas penghafalan Al-Qur'an. Analisis komprehensif mengindikasikan bahwa meskipun revolusi digital menciptakan berbagai rintangan bagi kontinuitas tradisi suci ini, inovasi dan adaptasi melalui utilisasi teknologi justru dapat menjadi kunci strategis dalam melestarikan keberlangsungan dan nilai spiritual dari praktik penghafalan Al-Qur'an di tengah arus transformasi zaman.

Implementasi nyata dari pemanfaatan teknologi dalam pembelajaran Al-Qur'an terwujud melalui pengembangan sistem pencarian ayat berbasis audio yang mampu membantu penghafal dalam mengidentifikasi dan memvalidasi bacaan secara otomatis. Inovasi terkini dalam ranah *automatic speech recognition* (ASR) untuk tilawah Al-Qur'an telah memperlihatkan progres yang menggembirakan dalam menangani kompleksitas identifikasi fonetik bahasa Arab. Khususnya, para akademisi telah mengadopsi metodologi *deep learning end-to-end*, dengan arsitektur CNN-Bidirectional GRU encoder yang dipadukan dengan CTC meraih pencapaian impresif yaitu 8,34% Word Error Rate (WER) dan 2,42% Character Error Rate (CER) dalam mengenali recitasi Al-Qur'an [5]. *Framework* komprehensif yang mengintegrasi pendekatan ASR mutakhir dengan prinsip-prinsip tajwid telah dibangun untuk menjamin ketepatan religius dalam sistem identifikasi [6].

Dalam spektrum yang lebih luas terkait pengolahan audio bahasa Arab, model yang berdasarkan *Wav2Vec2* telah mendemonstrasikan kemajuan yang signifikan, mencapai 24,3% WER dan 17,6% CER pada koleksi data bahasa Arab, yang mencerminkan reduksi 11,7% WER dibandingkan implementasi *Wav2Vec2* terdahulu [7]. Lebih lanjut, paradigma *self-supervised learning* canggih, meliputi *Wav2Vec2* dan *HuBERT*, telah sukses diaplikasikan untuk deteksi emosi dalam pelafalan bahasa Arab, mendemonstrasikan kapabilitas model-model tersebut untuk pengolahan audio Arab yang canggih [8]. Konstruksi dataset audio Arab spesifik seperti Aswat telah semakin mendukung *speech-representation learning* untuk implementasi ASR [9].

Kendati telah terjadi kemajuan substansial tersebut, masih terdapat celah fundamental dalam evaluasi perbandingan model *self-supervised learning* terdepan yang dikhususkan untuk tugas penelusuran ayat Al-Qur'an. Sementara riset yang tersedia telah mengonsentrasikan diri terutama pada akurasi speech recognition (metrik WER/CER), terjadi kekosongan yang signifikan dari penelitian yang secara metodis mengkomparasi model Data2Vec dan Wav2Vec dengan menggunakan parameter kesamaan semantik seperti cosine similarity untuk penelusuran ayat berbasis audio. Sistem yang ada saat ini lebih memprioritaskan akurasi transkripsi dibandingkan kapabilitas pencocokan semantik, yang menjadi aspek krusial untuk aplikasi praktis dimana pengguna memerlukan identifikasi ayat-ayat tertentu dari potongan audio.

Kesenjangan riset ini menjadi semakin menguatkan mengapa penelitian ini harus dilakukan. Karenanya, penelitian ini bertujuan mengisi gap tersebut melalui evaluasi komparatif Data2Vec dan Wav2Vec dengan menggunakan cosine similarity untuk aplikasi retrieval ayat Al-Qur'an, yang diharapkan dapat memberikan sumbangan berarti bagi pengembangan teknologi pendukung penghafalan Al-Qur'an yang lebih efisien dan adaptif terhadap kebutuhan pembelajaran masa kini.

2. Rumusan Masalah

Berdasarkan latar belakang di atas, maka rumusan masalah pada penelitian ini adalah:

1. Bagaimana kinerja model Data2Vec menggunakan *cosaine similarity*?
2. Bagaimana kinerja model wav2vec menggunakan *cosaine similarity*?
3. Menentukan model mana yang lebih cocok untuk digunakan dalam *retrival* ayat al quran berdasarkan cuplikan audio?

3. Batasan Masalah

Batasan masalah dari penelitian ini adalah sebagai berikut:

1. Penelitian dibatasi pada penggunaan model pretrained Data2Vec dan Wav2Vec2, tanpa pelatihan ulang (fine-tuning) pada data khusus Al-Qur'an
2. Eksperimen hanya dilakukan pada teks Al-Qur'an berbahasa Arab, terutama Surah-surah pendek seperti Al-Fatihah dan Juz Amma
3. Fokus sistem adalah pada pencarian kemiripan ayat, bukan pada koreksi tajwid, deteksi kesalahan pelafalan, atau penilaian kualitas tilawah
4. Evaluasi dilakukan menggunakan skenario recitation parsial atau tidak sempurna, bukan recitation lengkap

4. Tujuan Penelitian

Tujuan dari penelitian ini sebagai berikut:

1. Mengetahui kinerja model Data2Vec menggunakan *cosaine similarity*.
2. Mengetahui kinerja model wav2vec menggunakan *cosaine similarity*
3. Mengetahui perbandingan antara kinerja model Data2Vec dan wav2vec menggunakan *cosaine similarity*.

5. Manfaat Penelitian

Manfaat yang diharapkan dari hasil penelitian ini adalah sebagai berikut:

1. Memberikan pemahaman ilmiah yang lebih dalam tentang efektivitas model representasi audio seperti Data2Vec dan Wav2Vec2 dalam domain bahasa Arab, khususnya untuk tugas pencocokan kemiripan teks berbasis suara.
2. Memberikan acuan bagi peneliti dan pengembang aplikasi keislaman dalam memilih model *embedding* audio terbaik untuk diterapkan pada sistem pembelajaran Al-Qur'an berbasis suara

6. *The State of The Art*

Beberapa Penelitian telah dilakukan untuk membuat software maintainability yang baik dari masa ke masa. Dalam upaya mengembangkan sebuah aplikasi dengan metode yang telah dikembangkan maka dibutuhkannya proses studi literatur. Berikut merupakan penelitian yang sebelumnya telah dilakukan dengan metode yang serupa:

a. Alexei Baevski, dkk (2020). *Wav2vec 2.0* memperkenalkan kerangkan *self-supervised learning* untuk mempelajari representasi ucapan langsung dari *raw audio.* Arsitektur terdiri dari *convolutional feature encoder* untuk mengekstrasi fitur awal, *Transformer network* untuk memodelkan dependensi jangka panjang, dan *quantizatino module* untuk mengubah representasi menjadi kode diskrit yang digunakan dalam *contrastive learning.* Pendekatan ini secara signifikasn mengurangi kebutuhan data berlabel, menutup kesenjangan performa antara sistem ASR *fully-supervised* dan *low-resource*, serta mencapai *state-of-the-art* pada *benchmark* Librispeech 100h. Metode ini membuka arah baru pengenalan suara di bahasa dengan sumber daya terbatas melalui representasi ucapan yang lebih efisien secara data. [10]

b. Alexei Baevski, dkk (2022). Data2vec mengusulkan kerangka self-supervised learning multimodal yang seragam, mampu mempelajari representasi laten dari ucapan, teks dan citra menggunakan metode tunggal. Berbeda dengan pendekatan terdahulu yang memprediksi target spesifik modalitas, metode ini mempresdiksi representasi kontekstual laten dari input lengkap berdasarkan versi yang telah dilakukan *masking*. Proses pembelajaran dilakukan melalui mekanisme *self-distillation* dengan arsitekur Transformer, di mana model pelajar (*student*) mengestimasi keluaran model guru (*teacher*) yang dibekukan parameternya. Evaluasi

empiris menunjukkan bahwa data2vec mencapai kinerja *state-of-the-art*. [11]

c. Omar Mohamed & Salah A. Aly (2021). *Arabic Speech Emotion Recognition Employing Wav2vec2.0 and HuBERT Based on BAVED Dataset*. Makalah ini memperkenalkan model deep learning untuk pengenalan emosi dalam ucapan bahasa Arab menggunakan representasi audio canggih seperti wav2vec 2.0 dan HuBERT. Model-prinsip self-supervised dijalankan pada dataset tanpa label besar dan kemudian di-fine-tune pada dataset kecil (BAVED). Representasi fitur ini digunakan pada classifier berupa MLP dan Bi-LSTM. Hasil eksperimen menunjukkan bahwa wav2vec 2.0 memberikan performa terbaik dalam akurasi pengenalan emosi, konvergensi lebih cepat, dan stabilitas pelatihan dibanding HuBERT. Model ini mencapai akurasi hingga 89% pada dataset BAVED. Pendekatan ini menunjukkan efektivitas representasi self-supervised untuk tugas SER dalam bahasa Arab, terutama dengan data terbatas [12].

d. Yasser Shohoud, dkk (2023). memperkenalkan alat pencarian semantik untuk Al-Qur'an yang mendukung pencarian ayat berdasarkan tafsir yang sesuai dengan permintaan pengguna. Dengan melatih beberapa model pada kumpulan data besar yang terdiri dari lebih dari 30 tafsir—di mana masing-masing terhubung dengan satu ayat—metode ini mencari tensor tafsir yang memiliki kemiripan kosinus tertinggi dengan tensor representasi permintaan (prompt), lalu digunakan untuk mengindeks ayat yang relevan. Penerapan model SNxLM menghasilkan skor kosinus hingga 0,97, yang merefleksikan pencocokan tafsir Abdu untuk ayat yang terkait topik keuangan [13].

e. Matthijs Douze, dkk (2024). Pustaka C++ (dengan Python-*wrapper*) untuk pencarian kemiripan vektor, dilengkapi metode pengindeksan, kompresi, klasterisasi, dan transformasi. Fokus pada *trade-off* antara akurasi, kecepatan, memori, dengan optimisasi CPU/GPU dan antarmuka fleksibel. Menampilkan aplikasi di skala triliunan vektor, text retrieval, data mining, dan moderasi konten. Populer luas di komunitas—bukan untuk ekstraksi fitur atau manajemen transaksi [14].

No	Judul Jurnal dan Peneliti	Metode	Tujuan
1	wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations Peneliti: R Prajana, Prof.Kavitha S N (2021)	Wav2vec2, *contrastive masked prediction of quantized units*	Tujuan utama dari Wav2Vec 2.0 adalah memungkinkan pembelajaran representasi suara secara efektif tanpa membutuhkan dataset berlabel dalam jumlah besar. Penulis berupaya mempelajari representasi langsung dari sinyal audio mentah, mengurangi ketergantungan pada fitur buatan, serta menunjukkan bahwa pre-training self-supervised dapat secara signifikan meningkatkan kinerja automatic speech recognition (ASR) meskipun hanya menggunakan data berlabel yang terbatas. [10]
2	data2vec: A General Framework for Self-supervised Learning in Speech, Vision and Language Peneliti: Alexei Baevski, 2022	Data2vec, *self-distillation of contextual latent representations*	Data2vec bertujuan mengembangkan sebuah framework self-supervised yang seragam penerapannya untuk berbagai modalitas (gambar, suara, teks). Dalam pendekatannya, data2vec mengkombinasikan *masked prediction* dan *self-distillation* dengan rata-rata lapisan sebagai target untuk memprediksi representasi laten yang bersifat kontekstual

No	Judul Jurnal dan Peneliti	Metode	Tujuan
			dari seluruh input. Pendekatan ini telah terbukti memberikan performa state-of-the-art atau setara di benchmark-benchmark utama speech recognition, image classification, dan natural language understanding [11]
3.	Arabic Speech Emotion Recognition Employing Wav2vec2.0 and HuBERT Based on BAVED Dataset Penulis: Omar Mohamed, dkk (2021)	Menggunakan representasi kontekstual dari Wav2vec2 dan HuBERT. Setelah itu diklasifikasikan menggunakan MLP dan Bi-LSTM	Mengembangkan model pengenalan emosi dalam ucapan bahasa Arab dengan memanfaatkan representasi self-supervised. Validasi pada dataset BAVED menunjukkan model berbasis wav2vec 2.0 mencapai akurasi hingga 89%, mengungguli model HuBERT base (87%) dan HuBERT large (84%)
4.	Quranic Conversations: Developing a Semantic Search tool for the Quran using Arabic NLP Techniques Peneliti: Yasser Shohoud, dkk (2021)	Pelitian ini mengimplementasikan representasi teks berbasis *word embeddings* dan *transformer-based language models* (SNxLM, MPNet, MiniLM, DistilBERT, MultiFiT). Tafsir dalam bahasa Arab dan Inggris diproses dengan Word2Vec	Penelitian ini bertujuan merancang sistem pencarian semantik Al-Qur'an yang mampu menjawab pertanyaan konseptual ("What does the Qur'an say about ___?") secara kontekstual, dengan memanfaatkan tafsir sebagai jembatan interpretatif. Hasilnya menunjukkan bahwa model SNxLM menghasilkan kesesuaian semantik tertinggi, dengan skor kemiripan

No	Judul Jurnal dan Peneliti	Metode	Tujuan
		(CBOW) untuk menghasilkan vektor semantik. Pertanyaan pengguna dipetakan ke dalam ruang vektor, kemudian relevansi dihitung menggunakan *cosine similarity* untuk memperoleh ayat yang sesuai.	mencapai 0,97 pada topik tertentu, sehingga mendemonstrasikan potensi NLP untuk studi teks keagamaan.
5.	THE FAISS LIBRARY Peneliti: Matthijs Douze, dkk (2025)	Perancangan *toolkit* indexing: mencakup berbagai metode indexing seperti Inverted File (IVF), quantization, graph-based (HNSW), pra-pemrosesan (PCA), kompresi, serta optimisasi CPU/GPU dan antarmuka penyimpanan eksternal melalui abstraksi seperti InvertedLists dan InvertedListScanner.	Mendeskripsikan prinsip desain, trade-off antara akurasi dan efisiensi, serta antarmuka implementasi Faiss; memaparkan benchmark performa dan aplikasi nyata (triliun skala indeks, text retrieval, data mining, content moderation).

Tabel 1 *State of the art*

Pada Tabel 1 telah dipaparkan mengenai Kelima penelitian ini merepresentasikan kemajuan terkini dalam bidang pengenalan suara. Wav2Vec 2.0 memperkenalkan metode *self-supervised* untuk mengekstraksi suara langsung dari

sinyal audio mentah, mengurangi kebutuhan data berlabel besar, sementara data2vec memperluas pendekatan serupa menjadi *multimodal framework* yang seragam untuk suara, teks, dan gambar melalui mekanisme *self-distillation*. Kemudian, studi Arabic Speech Emotion Recognition menunjukkan penerapan praktis representasi kontekstual Wav2Vec 2.0 dan HuBERT dalam klasifikasi emosi bahasa Arab dengan hasil akurasi tinggi. Di ranah teks keagamaan, penelitian *Quranic Conversations* memanfaatkan *word embeddings* dan *transformer-based models* untuk membangun sistem pencarian semantik Al-Qur'an yang mampu menjawab pertanyaan konseptual secara kontekstual. Terakhir, FAISS *Library* berkontribusi dengan menyediakan infrastruktur indexing berskala besar yang efisien untuk pencarian vektor, melengkapi upaya-upaya sebelumnya dengan solusi komputasi yang mendukung penerapan nyata dalam pengembalian data berbasis *embedding*.

Kerangka pemikiran dalam penelitian tugas akhir ini akan dipaparkan pada Tabel 2 sebagai berikut:

Fenomena
Perkembangan teknologi digital memengaruhi metode tradisional hafalan Al-Qur'an. Namun, sistem yang ada lebih berfokus pada akurasi transkripsi (WER/CER) dan belum optimal untuk pencarian ayat berbasis audio dengan pendekatan semantik.
↓
Problem
Belum ada penelitian yang secara komprehensif membandingkan model self-supervised learning (Wav2Vec2 dan Data2Vec) untuk *audio-based retrieval* ayat Al-Qur'an menggunakan cosine similarity.
↓
Approach
Mengimplementasikan dua model modern (Wav2Vec2 dan Data2Vec) serta membandingkan performanya dengan metrik WER, CER, dan cosine similarity untuk retrieval ayat berbasis audio.
↓

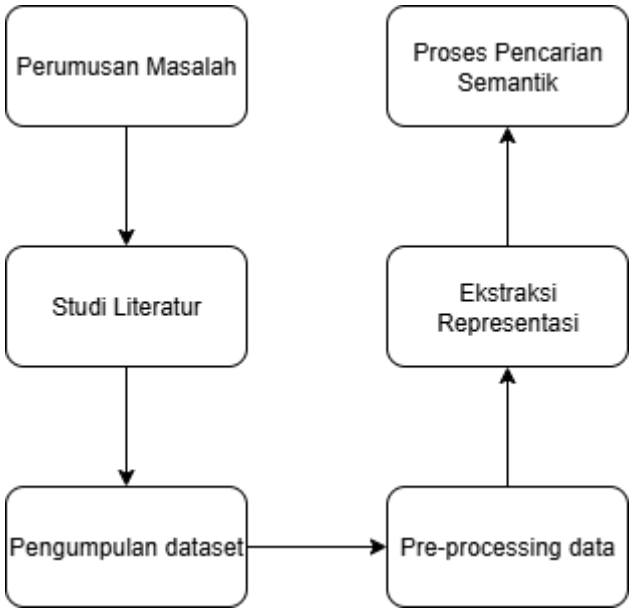
Dataset
Dataset audio Qur’an (misalnya Aswat), dilakukan preprocessing berupa normalisasi dan segmentasi ayat.

Output
Sistem pencarian ayat berbasis audio yang mendukung hafalan Al-Qur’an secara digital.

Tabel 2 Kerangka Pemikiran

7. Metode Penelitian

Penelitian ini dilakukan dengan beberapa tahap pengerjaan. Tahap-tahap tersebut di antara lain perumusan masalah, studi literatur, pengumpulan dan pengolahan dataset, impelementasi model, evaluasi performa, analisis hasil, serta penarikan kesimpulan. Alur tahapan tersebut dadpat dilihat pada Gambar 5.



Gambar 1 Metode Penelitian

1. Perumusan Masalah

Tahap awal penelitian ini diawali dengan identifikasi masalah yang berkembang dalam konteks pembelajaran Al-Qur'an berbasis digital. Penelitian terdahulu mayoritas berfokus pada pengukuran akurasi transkripsi dengan metrik Word Error Rate (WER) atau Character Error Rate (CER), namun belum secara memadai mengeksplorasi kemampuan model dalam pencocokan semantik ayat. Sementara itu, kebutuhan industri dan masyarakat saat ini tidak hanya terbatas pada deteksi kesalahan bacaan, melainkan juga pada identifikasi serta pencarian ayat secara otomatis berdasarkan kemiripan semantik. Hal ini menuntut adanya pendekatan yang mampu memanfaatkan representasi audio secara langsung untuk menghasilkan retrieval yang lebih akurat dan adaptif terhadap variasi pelafalan maupun keterbatasan data berlabel. Berdasarkan kesenjangan tersebut, penelitian ini dirancang untuk melakukan evaluasi komparatif antara model Wav2Vec2 dan Data2Vec dengan

memanfaatkan *cosine similarity*, sebagai upaya memberikan solusi yang lebih komprehensif dalam mendukung proses hafalan Al-Qur'an.

2. Studi literatur

Studi literatur merupakan suatu proses pencarian teori-teori atau penelitian terhadap sebuah masalah. Literatur yang digunakan dapat berupa buku, jurnal, artikel, ataupun *paper.* Kajian literatur dilakukan untuk memperoleh landasan teoretis yang relevan dan memperkuat kerangka konseptual penelitian. Literatur yang dianalisis mencakup:

A. Model **Wav2Vec2** sebagai pendekatan *self-supervised learning* berbasis *contrastive prediction* untuk representasi audio.

B. Model **Data2Vec** sebagai kerangka multimodal berbasis *self-distillation of contextual representations*.

C. Penelitian terdahulu mengenai penerapan **ASR pada Al-Qur'an**, termasuk arsitektur CNN-BiGRU dengan Connectionist Temporal Classification (CTC), serta studi yang mengimplementasikan HuBERT dan transformer.

D. Pendekatan pencarian semantik berbasis *vector embeddings* dengan pemanfaatan **FAISS** untuk *similarity search* dalam domain skala besar.

3. Pengumpulan dataset

Dataset yang digunakan mencakup audio bacaan Al-Qur'an yang telah tersedia secara terbuka dan legal, dengan variasi qari dan kondisi rekaman yang berbeda. Selain itu, teks mushaf standar ber-diakritik digunakan sebagai referensi. Keduanya berperan sebagai pasangan data yang akan dipetakan ke dalam ruang *embedding* semantik.

4. Pre-processing data

Pre-processing diperlukan untuk memastikan audio dan teks dalam kondisi yang sesuai sebelum diproses ekstrasi oleh model yang dipilih *pre-process* dapat berupa:

a. Segmentasi audio per ayat
b. Normalisasi sinyal dan pengurangan *noise*

c. Pembersihan teks dari simbol yang tidak relevan
d. Tokenisasi teks mushaf sebagai input untuk *embedding*

5. Ekstrasi Representasi

Pada tahap ini, data audio dan teks diproses untuk menghasilkan *embedding* dalam ruang vector. Poin yang mencakup ekstrasi representasi sebagai berikut:

a. Menggunakan model Wav2vec2 untuk mengekstrasi representasi audio berbasis *constrastive learning*
b. Menggunakan model Data2Vec untuk mengekstrasi representasi audio berbasis self-distillation
c. Menyediakan embedding teks mushaf standar untuk pembanding

6. Proses pencarian semantik

Tahap ini bertujuan untuk mencocokan representasi audio dengan representasi teks menggunakan pendekatan berbasis kesamaan vektor

e. Menghitung *cosine similarity* antara embedding audio dan embedding teks
f. Menentukan ayat dengan skor kesamaan tertinggi sebagai hasil pencarian
g. Menguji *robustnes* pencarian terhadap variasi bacaan, *noise* dan perbedaan qari

Lokasi Penelitian

Lokasi penelitian dapat dilakukan dimana saja dikarenakan penelitian tidak membutuhkan tempat khusus dalam pengambilan data maupun metode yang digunakan.

Jadwal Penelitian

NO	KEGIATAN	MINGGU												HASIL KESELURUHAN
		1	2	3	4	5	6	7	8	9	10	11	12	
1	Studi Literatur													Mengetahui bagaimana implementasi SOLID design principle dalam pengembangan sebuah perangkat lunak dan mengukurnya menggunakan C&K metrics
2	Pengumpulan dataset													Dataset siap digunakan (pembersihan, normalisasi, tokenisasi)
3	Pembuatan Embedding (Wav2vec2 & Data2Vec)													Vektor embedding untuk setiap ayat/bacaan Qur'an
4	Implementasi Sistem Pencarian Semantik (Cosine Similarity + FAISS)													Prototipe sistem pencarian berbasis embedding
5	Evaluasi Model													Hasil evaluasi performa model Wav2Vec2 vs Data2Vec
6	Penarikan Kesimpulan & penyusunan laporan													Kesimpulan akhir dan naskah laporan penelitian

DAFTAR PUSTAKA

[1] M. H. Jarrahi, D. L. Blyth, and C. Goray, “Mindful work and mindful technology: Redressing digital distraction in knowledge work,” *Digit. Bus.*, vol. 3, no. 1, p. 100051, June 2023, doi: 10.1016/j.digbus.2022.100051.

[2] Moh. A. Imam Sofii, “Menghafal Al Qur`an Di Era Digital: Problematis Dan Metodologis.,” *Al Furqan J. Ilmu Al Quran Dan Tafsir*, vol. 7, no. 1, pp. 1–17, June 2024, doi: 10.58518/alfurqon.v7i1.2436.

[3] O. Shaklawoon, A. Shafter, M. Abuzaraida, A. Zeki, and Z. Mahmod, *Monitoring the memorization of the Holy Qur’an based on Speech Recognition and NLP Techniques*. 2023.

[4] M. Mutathahirin, A. Jaafar, and N. R. Kamaruzaman, “A Systematic Literature Review (SLR) on Quranic Memorization: Benefits, Methods, and Innovations,” vol. 6, no. 2.

[5] A. A. Harere and K. A. Jallad, “Quran Recitation Recognition using End-to-End Deep Learning,” May 10, 2023, *arXiv*: arXiv:2305.07034. doi: 10.48550/arXiv.2305.07034.

[6] S. Al-Fadhli, H. Al-Harbi, and A. Cherif, “Speech Recognition Models for Holy Quran Recitation Based on Modern Approaches and Tajweed Rules: A Comprehensive Overview,” *Int. J. Adv. Comput. Sci. Appl.*, vol. 14, no. 12, 2023, doi: 10.14569/IJACSA.2023.0141297.

[7] N. Oukas, T. Zerrouki, S. Haboussi, and H. Djettou, “Arabic Speech Recognition Using Deep Learning and Common Voice Dataset,” in *2022 International Conference on Innovation and Intelligence for Informatics, Computing, and Technologies (3ICT)*, Sakheer, Bahrain: IEEE, Nov. 2022, pp. 642–647. doi: 10.1109/3ICT56508.2022.9990834.

[8] O. Mohamed and S. A. Aly, “Arabic Speech Emotion Recognition Employing Wav2vec2.0 and HuBERT Based on BAVED Dataset,” 2021, *arXiv*. doi: 10.48550/ARXIV.2110.04425.

[9] L. Alkanhal, A. Alessa, E. Almahmoud, and R. Alaqil, “Aswat: Arabic Audio Dataset for Automatic Speech Recognition Using Speech-Representation Learning,” in *Proceedings of ArabicNLP 2023*, Singapore (Hybrid): Association for Computational Linguistics, 2023, pp. 120–127. doi: 10.18653/v1/2023.arabicnlp-1.10.

[10] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, “wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations,” Oct. 22, 2020, *arXiv*: arXiv:2006.11477. doi: 10.48550/arXiv.2006.11477.

[11] A. Baevski, W.-N. Hsu, Q. Xu, A. Babu, J. Gu, and M. Auli, “data2vec: A General Framework for Self-supervised Learning in Speech, Vision and Language,” Oct. 25, 2022, *arXiv*: arXiv:2202.03555. doi: 10.48550/arXiv.2202.03555.

[12] O. Mohamed and S. A. Aly, “Arabic Speech Emotion Recognition Employing Wav2vec2.0 and HuBERT Based on BAVED Dataset,” Oct. 09, 2021, *arXiv*: arXiv:2110.04425. doi: 10.48550/arXiv.2110.04425.

[13] Y. Shohoud, M. Shoman, and S. Abdelazim, “Quranic Conversations: Developing a Semantic Search tool for the Quran using Arabic NLP Techniques,” Nov. 09, 2023, *arXiv*: arXiv:2311.05120. doi: 10.48550/arXiv.2311.05120.

[14] M. Douze *et al.*, “The Faiss library,” Feb. 11, 2025, *arXiv*: arXiv:2401.08281. doi: 10.48550/arXiv.2401.08281.

[15] P. Primus and G. Widmer, “Fusing Audio and Metadata Embeddings Improves Language-based Audio Retrieval,” July 02, 2024, *arXiv*: arXiv:2406.15897. doi: 10.48550/arXiv.2406.15897.