IDENTIFIKASI PENYAKIT PARU-PARU BERDASARKAN SUARA DAN RIWAYAT PASIEN MENGGUNAKAN MODEL *CROSS-ATTENTION VIDEO VISION TRANSFORMER*

NASKAH SKRIPSI

Husni Na'fa Mubarok
121450078

PROGRAM STUDI SAINS DATA
FAKULTAS SAINS
INSTITUT TEKNOLOGI SUMATERA
LAMPUNG SELATAN

2024

IDENTIFIKASI PENYAKIT PARU-PARU BERDASARKAN SUARA DAN RIWAYAT PASIEN MENGGUNAKAN MODEL *CROSS-ATTENTION VIDEO VISION TRANSFORMER*

NASKAH SKRIPSI
Diajukan sebagai syarat maju seminar hasil

Husni Na'fa Mubarok
121450078

PROGRAM STUDI SAINS DATA
FAKULTAS SAINS
INSTITUT TEKNOLOGI SUMATERA
LAMPUNG SELATAN

2024

HALAMAN PENGESAHAN

Naskah Tugas Akhir untuk Seminar Hasil dengan judul "**Identify lung disease based on sound and patient history using the *Cross-Attention Video Vision Transformer* model**" adalah benar dibuat oleh saya sendiri dan belum pernah dibuat dan diserahkan sebelumnya, baik sebagian ataupun seluruhnya, baik oleh saya ataupun orang lain, baik di Institut Teknologi Sumatera maupun di institusi pendidikan lainnya.

Lampung Selatan, 06 April 2024

Penulis,

Husni Na'fa Mubarok
NIM. 121450078

Diperiksa dan disetujui oleh,

Pembimbing I	Pembimbing II
Christyan Tamaro Nadeak, M.Si	**Luluk Muthoharoh, M.Si**
NRK. 1993120420211415	**NIP. 199504112022032014**

Disahkan oleh,

Koordinator Program Studi Sains Data
Fakultas Sains
Institut Teknologi Sumatera

Tirta Setiawan, S.Pd., M.Si
NIP. 199008222022031003

Sidang Tugas Akhir :

Penguji I : Mika Alvionita S, M.Si
Penguji II : Penguji 2

HALAMAN PERNYATAAN ORISINALITAS

Skripsi ini adalah karya saya sendiri dan semua sumber baik yang dikutip maupun yang dirujuk telah saya nyatakan benar.

Nama	:	**Husni Na'fa Mubarok**
NIM	:	**121450078**
Tanda tangan	:	
Tanggal	:	**06 April 2024**

HALAMAN PERNYATAAN PERSETUJUAN PUBLIKASI UNTUK KEPENTINGAN AKADEMIS

Sebagai civitas akademik Institut Teknologi Sumatera, saya yang bertanda tangan di bawah ini:

Nama	:	Husni Na'fa Mubarok
NIM	:	121450078
Program Studi	:	Sains Data
Fakultas	:	Sains
Jenis karya	:	Skripsi

demi pengembangan ilmu pengetahuan, menyetujui untuk memberikan Hak Bebas Royalti Noneksklusif (*Non-Exclusive Royalty Free Right*) kepada Institut Teknologi Sumatera atas karya ilmiah saya yang berjudul:

Identifikasi penyakit paru-paru berdasarkan suara dan riwayat pasien menggunakan model *Cross-Attention Video Vision Transformer*

beserta perangkat yang ada (jika diperlukan). Dengan Hak Bebas Royalti Noneksklusif ini Institut Teknologi Sumatera berhak menyimpan, mengalihmedia/format-kan, mengelola dalam bentuk pangkalan data (*database*), merawat, dan memublikasikan tugas akhir saya selama tetap mencantumkan nama saya sebagai penulis/pencipta dan sebagai pemilik Hak Cipta.

Demikian pernyataan ini saya buat dengan sebenarnya.

Dibuat di	:	Lampung Selatan
Pada tanggal	:	06 April 2024

Yang menyatakan

Husni Na'fa Mubarok

ABSTRAK

Identifikasi penyakit paru-paru berdasarkan suara dan riwayat pasien menggunakan model *Cross-Attention Video Vision Transformer*

Husni Na'fa Mubarok (121450078)
Pembimbing I: Christyan Tamaro Nadeak, M.Si
Pembimbing II: Luluk Muthoharoh, M.Si

Lorem ipsum dolor sit amet, consectetur adipisicing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum. Sed ut perspiciatis unde omnis iste natus error sit voluptatem accusantium doloremque laudantium, totam rem aperiam, eaque ipsa quae ab illo inventore veritatis et quasi architecto beatae vitae dicta sunt explicabo. Nemo enim ipsam voluptatem quia voluptas sit aspernatur aut odit aut fugit, sed quia consequuntur magni dolores eos qui ratione voluptatem sequi nesciunt.

Kata kunci: ini, itu, ini, itu

ABSTRACT

Identify lung disease based on sound and patient history using the Cross-Attention Video Vision Transformer model

Husni Na'fa Mubarok (121450078)
Advisor I : Christyan Tamaro Nadeak, M.Si
Advisor II: Luluk Muthoharoh, M.Si

Lorem ipsum dolor sit amet, consectetur adipisicing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum. Sed ut perspiciatis unde omnis iste natus error sit voluptatem accusantium doloremque laudantium, totam rem aperiam, eaque ipsa quae ab illo inventore veritatis et quasi architecto beatae vitae dicta sunt explicabo. Nemo enim ipsam voluptatem quia voluptas sit aspernatur aut odit aut fugit, sed quia consequuntur magni dolores eos qui ratione voluptatem sequi nesciunt.

Keywords : *this, that, this, thatthis*.

MOTTO

Ini mottoku, mana motto-mu?.

HALAMAN PERSEMBAHAN

Untuk Emak dan Bapak
di kampung

KATA PENGANTAR

Puji syukur penulis panjatkan ke hadirat Allah SWT atas berkah dan rahmat-Nya sehingga skripsi ini dapat terselesaikan dengan baik. Skripsi ini dibuat untuk menyelesaikan pendidikan jenjang sarjana pada Institut Teknologi Sumatera. Penyusunan skripsi ini banyak mendapat bantuan dan dukungan dari berbagai pihak sehingga dalam kesempatan ini, dengan penuh kerendahan hati, penulis mengucapkan terima kasih kepada:

1. Prof. Xxxx Xxxx selaku Rektor Institut Teknologi Sumatera,
2. Prof. Yyyy Yyyy selaku Dekan Fakultas Sains Institut Teknologi Sumatera,
3. Dr. Zzzz Zzzz selaku Koordinator Program Studi,
4. Prof. Dr. Nama selaku dosen pembimbing pertama yang telah membimbing,
5. Nama , S.Si., M.Si. selaku dosen pembimbing kedua yang selalu membantu, dan
6. Cantumkan pihak-pihak lain yang membantu penelitian tugas akhir, termasuk sumber data, tempat riset, rekan satu TA, dan-lain-lain.

Penulis menyadari bahwa penyusunan Skripsi ini jauh dari sempurna. Akhir kata penulis mohon maaf yang sebesar-besarnya apabila ada kekeliruan di dalam penulisan skripsi ini.

Lampung Selatan, 06 April 2024

Husni Na'fa Mubarok

DAFTAR ISI

DAFTAR GAMBAR

DAFTAR TABEL

BAB I

PENDAHULUAN

1.1 Latar Belakang

Chronic obstructive pulmonary disease (COPD) adalah gangguan pernapasan yang umum dan progresif yang ditandai dengan keterbatasan aliran udara yang terus-menerus dan sering dikaitkan dengan penyakit paru-paru lainnya seperti bronkitis kronis dan asma [1]. Menurut data *World Health Organization* (WHO) *Chronic obstructive pulmonary disease* (COPD) merupakan penyebab kematian keempat di seluruh dunia, menyebabkan 3.5 juta kematian pada tahun 2021 [2], dan diperkirakan COPD akan menjadi penyebab kematian ketiga di dunia pada tahun 2030 [3]. Berdasarkan survei data BPJS tahun 2024, hampir 19 juta jumlah pasien COPD yang berobat ke rumah sakit dengan penyakit tersebut [4]. Oleh sebab itu, diperlukan metode yang dapat mengidentifikasi jenis penyakit paru-paru berdasarkan gejala yang diketahui salah satunya yaitu suara.

Suara batuk atau paru-paru mengandung banyak informasi tentang kondisi paru-paru dan dapat digunakan untuk menilai serta mendiagnosis penyakit pernapasan kronis [5]. Penggunaan suara untuk identifikasi penyakit paru meningkatkan minat terhadap perawatan medis tanpa kontak untuk pemeriksaan paru-paru secara otomatis. Model Deep Learning banyak digunakan peneliti dalam menganalisis suara seperti LungRN+NL [6] yang menggunakan augmentasi data campuran dan arsitektur ResNet [7] untuk mengatasi ketidakseimbangan kelas data. RespireNet [8] menggunakan model *pre-trainer* pada ImageNet dengan strategi *fine-tuning device-specific*.

Model *General-Purpose* representasi audio lainnya seperti CLAP [9] menggunakan dua encoder untuk memproses input yaitu *Audio Encoder* yang memproses input suara dan *Text Encoder* yang memproses input berupa teks. Namun, penggunaan dua encoder mengakibatkan beban komputasi yang tinggi sehingga memerlukan sumber daya komputasi yang besar.

Oleh sebab itu, pada penelitian ini mengadopsi konsep dari *Video Vision Transformer* [10] yang menggunakan embedding spasial Z_s dan temporal Z_t dengan input berupa representasi MFCC dan fitur riwayat pasien pada satu encoder yang sama sehingga mengurangi beban komputasi saat proses *training*. *Cross-Attention* [11] digunakan agar suatu *sequence* dapat memperhatikan

informasi dari *sequence* lainnya.

Secara keseluruhan, kontribusi penelitian ini dapat dirangkum yaitu MFCC digunakan untuk merepresentasikan spektrum daya jangka pendek dari suatu suara yang membantu model memahami dan memproses suara manusia secara lebih efektif. Penelitian ini mengadopsi model *Video Vision Transformer* untuk memahami fitur temporal dan spasial dari data audio. Data riwayat pasien digunakan untuk memperkaya fitur tanpa menambah encoder fitur sehingga beban komputasi menjadi lebih ringan.

1.2 Rumusan Masalah

Bagian ini menjadi salah satu bagian penting dalam Pendahuluan. Setelah paparan Latar Belakang 1.1, maka masalah yang diangkat pada pekerjaan penelitian perlu dirumuskan dengan baik. Pertanyaan apa yang akan dijawab dalam penelitian dapat ditulis dalam kalimat tanya ataupun tidak.

Berdasarkan latar belakang yang telah dijelaskan sebelumnya, berikut merupakan rumusan masalah pada penelitian tugas akhir ini:

1. Bagaimana penerapan konsep *Video Vision Transformer* pada data suara penyakit paru dengan penambahan fitur riwayat pasien menggunakan metode *Cross-Attention* dapat mengidentifikasi jenis penyakit paru-paru?
2. Seberapa baik penerapan konsep *Video Vision Transformer* dalam mengklasifikasikan jenis penyakit paru-paru berdasarkan data suara?

1.3 Tujuan Penelitian

Eros reprimique vim no. Alii legendos volutpat in sed, sit enim nemore labores no. No odio decore causae has. Vim te falli libris neglegentur, eam in tempor delectus dignissim, nam hinc dictas an.

Tujuan dari penelitian ini berdasarkan rumusan masalah yang juga menjadi dasar dilakukannya penelitian ini adalah sebagai berikut:

1. Membuat model *Deep Learning* untuk mengklasifikasikan jenis penyakit paru-paru menggunakan konsep Model *Vision Transformer* dengan metode *Cross-Attention*.
2. Mengevaluasi performa model Transformer dalam mengidentifikasi fitur suara dan riwayat pasien sehingga menghasilkan klasifikasi yang sesuai

1.4 Batasan Masalah

1. Penelitian ini hanya menggunakan data suara batuk dan riwayat pasien tanpa menyertakan identitas atau informasi lainnya.
2. Jenis penyakit paru-paru yang di identifikasi terbatas pada dataset yang digunakan.

BAB II

TINJAUAN PUSTAKA

2.1 Penelitian Terdahulu

Tabel 2.1 Penelitian Terdahulu

Author(s)	Metode	Hasil Penelitian
Hao Xue et al.[12]	Transformer-CP	Model berbasis Transformer dengan Constrantive Pre-training yang menggunakan random masking dengan fitur ekstraksi MFCC mendapatkan akurasi 87,74% pada data suara batuk pasien covid-19
Li Xiao et al.[13]	LungAdapter	Metode ini menggabungkan blok *trainable* ke dalam model AST yang telah dilatih sebelumnya, yang memungkinkan ekstraksi informasi penting tentang klasifikasi suara paru-paru dari model. Model mencapai kinerja yang baik dengan score 62,40%.
Victor Basu et al.[14]	GRU, MFCC	Lapisan GRU (gated recurrent unit) digunakan untuk memecahkan masalah gradien yang hilang dalam RNN standar. Akurasi: 95,67 ± 0,77%.

2.2 Penyakit Paru

2.2.1 *Chronic obstructive pulmonary disease* (COPD)

Chronic Obstructive Pulmonary Disease (COPD) memiliki berbagai gejala yang sering kali dapat disalahartikan sebagai kondisi pernapasan lainnya, sehingga mempersulit diagnosis dan penanganannya. Gejala utamanya meliputi batuk kronis [15], dispnea [16], dan produksi sputum [17], yang tumpang tindih dengan kondisi seperti asma dan bronkitis.

2.2.2 Asma

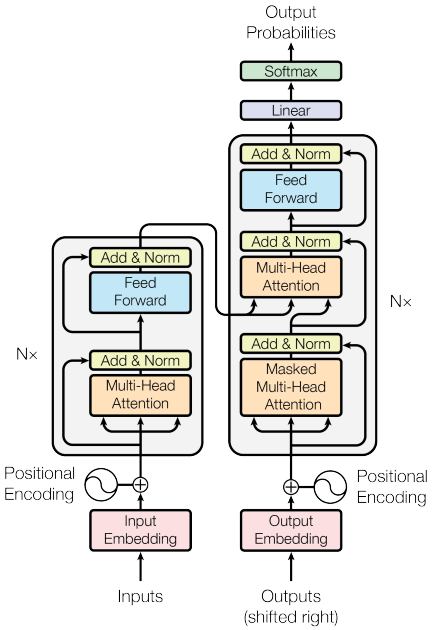
Asma adalah suatu kondisi pernapasan kronis yang ditandai dengan gejala yang bervariasi, terutama mengi, batuk, dada terasa sesak, dan dispnea yang intensitas dan frekuensinya dapat berfluktuasi [18]. Gejala-gejala ini sering kali timbul dari pemicu seperti alergen, infeksi, atau olahraga, dan mungkin terjadi secara intermiten, dengan beberapa pasien mengalami periode bebas gejala [18], [19].

2.3 *Mel Frequency Cepstral Coefficient* (MFCC)

MFCC memanfaatkan persepsi frekuensi suara telinga manusia, menggunakan skala Mel non-linier untuk mengubah sinyal audio menjadi representasi yang lebih relevan secara persepsi. Transformasi ini dicapai melalui serangkaian langkah, termasuk *pre-emphasis*, *windowing* dan *framing*, transformasi Fourier, pemrosesan *Mel filter bank*, dan analisis *cepstral*[20], [21].

2.4 Transformer

Arsitektur Transformer pertama kali diperkenalkan oleh Vaswani et al. (2017) pada paper yang berjudul "*Attention Is All You Need*"[22] yang terdiri dari dua bagian utama yaitu encoder dan decoder. Karena fleksibilitasnya dalam menangani *sequence* dan kemampuannya dalam memahami konteks, peneliti banyak mengadopsi arsitektur ini pada berbagai media seperti suara [13], gambar [23] dan video [10].



Gambar 2.1 Arsitektur Transformer

2.4.1 *Positional Encoding*

Arsitektur Transformer bersifat paralel sehingga diperlukan *Positional Encoding* untuk memberikan informasi urutan pada *sequence*. *Positional Encoding* memiliki ukuran dimensi yang sama dengan *embedding* input sehingga dapat dijumlahkan. Positional encoding dijelaskan pada persamaan 2.1 dan 2.2.

$$PE_{(pos,2i)} = \sin\left(pos/10000^{2i/d_{\text{model}}}\right) \tag{2.1}$$

$$PE_{(pos,2i+1)} = \cos\left(pos/10000^{2i/d_{\text{model}}}\right) \tag{2.2}$$

Keterangan:

- $PE_{(pos,2i)}$ adalah posisi pada indeks ganjil
- $PE_{(pos,2i+1)}$ adalah posisi pada indeks genap
- pos adalah posisi sampel/data dalan *sequence*
- i adalah dimensi data
- d_{model} adalah ukuran embedding

pos merupakan posisi dan i merupakan dimensi. $2i$ dan $2i + 1$ adalah indeks ganjil dan genap dalam embedding. Positional encoding sinusoidal dipilih karena memungkinkan model untuk mengekstrapolasi *sequnece* yang lebih panjang daripada yang ditemui selama pelatihan.

2.4.2 *Attention Mechanism*

Fungsi *Attention* dapat dideskripsikan sebagai pemetaan *Query* dan sekumpulan pasangan *Key-Value* ke suatu *Output*, di mana *Query*, *Key*, *Value*, dan *Output* semuanya merupakan vektor. *Self-Attention* menghubungkan semua posisi dengan jumlah operasi *sequence* yang konstan sehingga mempercepat operasi dibandingkan layer recurrent pada RNN. Fungsi *Attention* dijelaskan pada persamaan 2.3

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{Head\_DIM}}\right)V \tag{2.3}$$

Keterangan:

- Attention(Q, K, V) adalah mekanisme *self-attention*
- Q adalah *Query*
- K adalah *Key*
- V adalah *Value*
- $Head\_DIM$ adalah dimensi embedding

Multi-Head Attention memungkinkan model untuk secara bersamaan memberikan *Attention* informasi dari subruang representasi yang berbeda pada posisi yang berbeda 2.4.

$$\begin{aligned} \text{MultiHead}(Q, K, V) &= \text{concat}(\text{head}_1, \ldots, \text{head}_h)W^O \\ \text{dengan head}_i &= \text{Attention}(QW_i^Q, KW_i^K, VW_i^V) \end{aligned} \tag{2.4}$$

Keterangan:

- MultiHead(Q, K, V) adalah mekanisme multi-head attention
- head$_i$ adalah mekanisme attention
- W_i^Q adalah bobot untuk *Query*
- W_i^K adalah bobot untuk *Key*
- W_i^V adalah bobot untuk *Value*
- W^O adalah bobot untuk multi-head attention

2.4.3 *Position-wise Feed-Forward Networks*

Masing-masing lapisan dalam encoder dan decoder terdapat *Feed-Forward Networks* yang saling terhubung, yang diterapkan ke setiap posisi secara terpisah dan identik. Fungsi *Feed-Forward Networks* dapat dilihat pada persamaan 2.5.

$$\text{FFN}(x) = \max(0, xW_1 + b_1)W_2 + b_2 \tag{2.5}$$

Keterangan:

- FFN(x) adalah mekanisme Feed-Forward Networks
- x adalah nilai input neuron
- W_1 adalah bobot pertama
- b_1 adalah bias pertama
- W_2 adalah bobot kedua
- b_2 adalah bobot kedua

Meskipun transformasi linier sama di berbagai posisi, FNN menggunakan parameter yang berbeda pada setiap lapisan.

2.4.4 *Layer Nornalization*

Layer Normalization digunakan untuk menormalkan output dari setiap jaringan, sehingga distribusi nilainya stabil selama *training* 2.6. Parameter γ dan β merupakan parameter pelatihan dan ϵ merupakan nilai konstan 10^{-6}.

$$\text{LayerNorm}(x) = \frac{x - \mu}{\sqrt{\sigma^2 + \epsilon}} \cdot \gamma + \beta \tag{2.6}$$

Keterangan:

- LayerNorm(x) adalah Layer Normalisasi
- x adalah nilai input layer
- μ adalah rataan

- σ^2 adalah standar deviasi
- ϵ adalah nilai galat agar pembagi tidak bernilai 0
- γ dan β adalah parameter pelatihan

2.5 *Video Vision Transformer*

Pada ViViT, video input V dengan dimensi D (depth), H (tinggi), W (lebar), dan C (channel) dibagi-bagi menjadi potongan-potongan kecil yang disebut *patch*[10]. Setiap *patch* P memiliki dimensi $P_d \times P_h \times P_w \times C$. Proses pembagian video menjadi *patch* ini dapat divisualisasikan sebagai persamaan 2.7

$$V \in \mathbb{R}^{D \times H \times W \times C} \rightarrow \text{Reshape} \rightarrow P \in \mathbb{R}^{N \times (P_d \times P_h \times P_w \times C)} \tag{2.7}$$

Keterangan:

- V adalah input embedding video
- P adalah input patch embedding
- Reshape adalah mengubah input video ke patch
- D adalah ukuran *depth* kedalaman/frame
- H adalah ukuran *height* atau tinggi
- W adalah ukuran *width* atau lebar
- C adalah jumlah *channel* atau kanal

di mana $N = (D/P_d) \times (H/P_h) \times (W/P_w)$ adalah jumlah total *patch* yang dihasilkan. Selanjutnya, setiap *patch* P diproyeksikan ke dalam ruang *embedding* dengan dimensi d melalui sebuah lapisan linear. Hasil proyeksi ini membentuk sebuah *sequence* token Z dengan dimensi $N \times d$ sesuai persamaan 2.8

$$Z = \text{Linear}(P) \in \mathbb{R}^{N \times d} \tag{2.8}$$

Keterangan:

- Z adalah hasil proyeksi *patch embedding*
- P adalah *patch embedding*

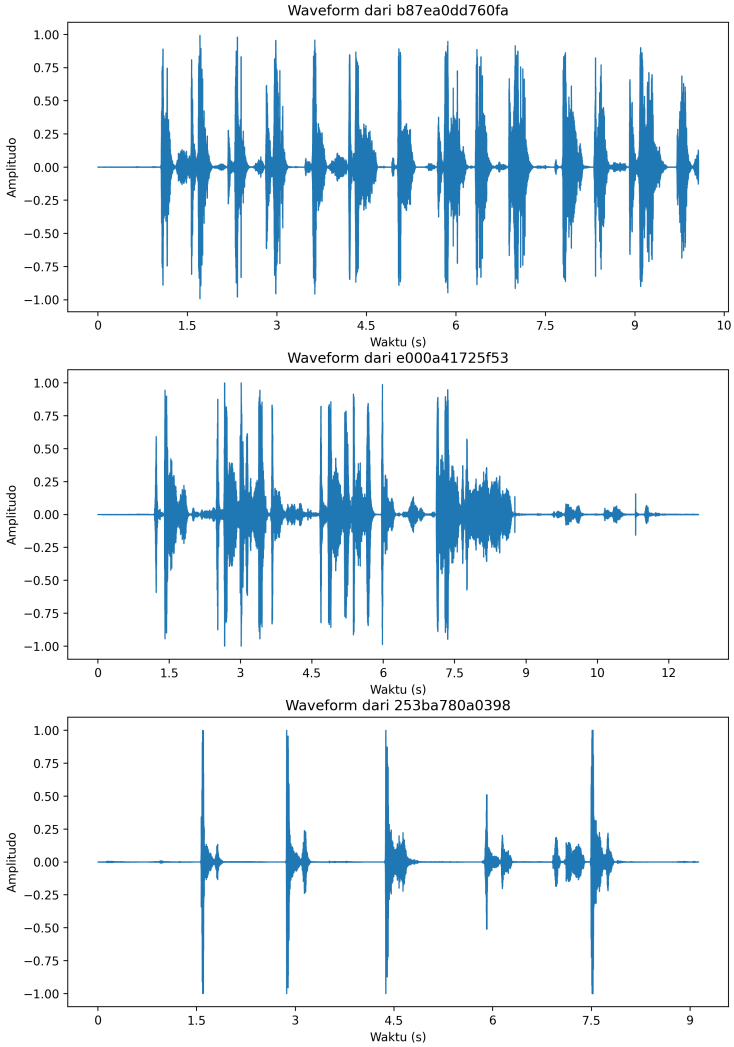
Sequence token Z ditambahkan dengan *positional encoding* yang kemudian menjadi input untuk *encoder transformer*. Terdapat dua metode utama untuk menghasilkan *patch embedding* yaitu, *Uniform Frame Sampling* yang hanya memiliki informasi spasial dan *Tubelet Embedding* yang menghasilkan *patch* yang mencakup informasi spasial dan temporal.

BAB III

METODE PENELITIAN

3.1 Deskripsi Data

Dataset yang digunakan adalah *Medical Sound Classification Challenge* yang diselenggarakan oleh Himanshu Kaushik pada platform Kaggle[24]. Dataset ini memiliki 882 jumlah data yang terdiri dari tiga jenis data untuk setiap individu yaitu audio batuk, vowel dan fitur riwayat pasien. Data diidentifikasi menggunakan candidateID yang ditetapkan untuk setiap orang. File suara dan embedding terdapat di folder candidateID yang diberikan. Dataset terbagi menjadi dua jenis yaitu 544 untuk data latih dan 338 untuk data uji yang belum memiliki label. Label penyakit terdiri dari 3 kelas yaitu asma, COPD dan pasien sehat.

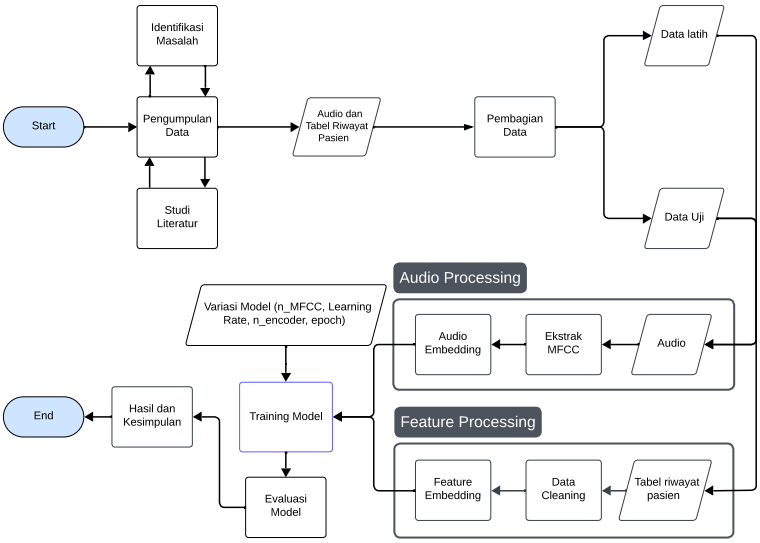


Gambar 3.1 Visualisasi gelombang suara tiap kelas

Tabel 3.1 Tabel riwayat pasien

candidateID	age	gender	tbContact	wheezing	phlegmCough	familyAsthma	feverHistory	coldPresent	packYears	disease
2bbd6c5ecf1ce	55	1	0.0	0.0	0.0	1.0	0.0	0.0	0	1
75fa6e335b5ca	65	0	0.0	1.0	0.0	0.0	0.0	1.0	560	2
7dc99cfcb5aa	43	0	0.0	1.0	0.0	0.0	0.0	0.0	0	1
59cf4a7821471	74	0	0.0	1.0	0.0	0.0	0.0	0.0	800	2
59f9fe56c2f12	28	0	0.0	0.0	1.0	0.0	0.0	0.0	0	0
caee891a86d8d	56	1	0.0	1.0	0.0	0.0	0.0	0.0	0	1
5b8578b39385f	31	0	0.0	0.0	1.0	0.0	0.0	0.0	0	0
f3e7d50ce7288	35	0	0.0	1.0	1.0	0.0	0.0	0.0	0	1
ad5fa122d4efb	56	1	0.0	0.0	0.0	1.0	0.0	0.0	0	0
14b58d18c66c7	57	0	0.0	1.0	1.0	0.0	0.0	0.0	0	0
7959121db060d	48	0	0.0	1.0	0.0	0.0	0.0	0.0	0	1
77aa6a34f6da1	21	0	0.0	0.0	1.0	1.0	0.0	0.0	0	1
069276518f6f	25	1	0.0	0.0	0.0	1.0	0.0	0.0	0	2
3d71862ec1801	48	1	0.0	0.0	1.0	1.0	0.0	0.0	0	1
1868d92d2db4	20	0	0.0	0.0	0.0	1.0	0.0	1.0	360	2
fbf8398bd0136	70	0	0.0	0.0	1.0	0.0	0.0	0.0	0	1
0343c366074ec	25	0	0.0	1.0	1.0	0.0	0.0	0.0	0	1
...	...	...	...	...	...	...	...	...	...	...
e17cb00bd9677	66	0	0.0	0.0	1.0	0.0	0.0	1.0	0	2

3.2 Rancangan Penelitian



Gambar 3.2 *Flowchart* rancangan penelitian

3.2.1 Pengumpulan Data

Data diunduh ke dalam *notebook* menggunakan kaggle-api dengan fungsi yang telah disediakan di laman kompetisi. File unduhan berupa .zip yang harus diekstrak terlebih dahulu.

3.2.2 Pembagian Data

Data dibagi menjadi data latih dan data uji dengan perbandingan 80% untuk data latih dan 20% untuk data uji. Digunakan parameter random state agar proses pembagian data bisa direproduksi ulang. Dilakukan *stratified sampling* berdasarkan kolom target yaitu *disease* untuk memastikan distribusi kelas tetap sama di kedua set.

3.2.3 Pemrosesan Data Tabel Riwayat Pasien

Dilakukan *preprocessing* seperti pengecekan nilai kosong, penghapusan fitur yang tidak relevan untuk memastikan kualitas data yang baik agar model dapat menangkap informasi yang relevan pada data.

3.2.4 Pemrosesan Suara

Data suara yang digunakan pada penelitian ini berupa suara batuk *cough.wav* pasien yang berisi minimal 3 kali batuk, dengan durasi rekaman maksimal 15 detik. Dilakukan ekstraksi fitur menggunakan *Mel frequency Capstral Coefficients*

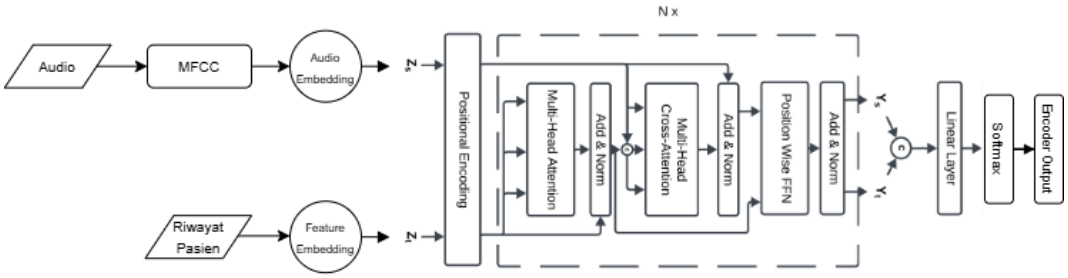
(MFCC), yaitu dengan cara memfilter secara logaritmik pada frekuensi di atas 1000 Hz dan secara linier pada frekuensi di bawah 1000 Hz. MFCC dapat meningkatkan sensitivitas pada suara dengan frekuensi rendah dan sebaliknya, pada suara dengan frekuensi tinggi MFCC dapat mengurangi sensitivitas dalam menangkap suara [25].

3.2.5 Pemodelan

Pada tahapan ini dilakukan rancangan model yang akan digunakan, seperti desain arsitektur, *loss function*, dan *optimizer* beserta *hyperparameter*-nya. Pada penelitian ini juga akan menggunakan beberapa desain arsitektur yang berfokus pada kedalaman seperti jumlah *block* dan *hyperparameter* jumlah *attention head*.

3.2.6 Arsitektur

Penelitian ini akan menggunakan MFCC untuk mengekstraksi fitur audio dan serta *embedding* untuk data riwayat pasien.



Gambar 3.3 Desain rancangan arsitektur model.

Input berupa audio dengan fitur MFCC berukuran $T \times F$ (dengan T adalah jumlah frame dan F adalah jumlah koefisien MFCC) dan data riwayat pasien berukuran $N \times d_{feat}$ (N adalah jumlah data riwayat, d_{feat} adalah dimensi fitur) akan diubah menjadi embedding berukuran $T \times d_{embed}$ dan $N \times d_{embed}$ melalui *embedding layer*. *Embedding* ini ditambahkan dengan *positional encoding* dan diteruskan ke *encoder* berbasis transformer, di mana mekanisme *self-attention* dan *cross-attention* digunakan untuk menangkap hubungan antar fitur audio serta interaksi antara audio dan riwayat pasien. Setelah melalui *feedforward network* (FFN), hasil akhir *encoder* berupa *embedding multimodal* dikombinasikan dan diproyeksikan menggunakan linear layer, diikuti oleh *softmax* untuk menghasilkan output akhir dari token [cls] berupa representasi yang dapat digunakan untuk klasifikasi.

3.2.7 *Loss Function*

Fungsi kerugian yang digunakan *Categorical Cross-Entropy* untuk setiap kelas 3.1:

$$L(y, \hat{y}) = -\sum_{i=1}^{C} y_i \log(\hat{y}_i) \tag{3.1}$$

Keterangan:

- $L(y, \hat{y})$ adalah nilai *Categorical Cross-Entropy*
- y adalah nilai target asli
- $\hat{y}_i$ adalah nilai prediksi

Di mana $L(y, \hat{y})$ merupakan *categorical cross-entropy loss*. y_i adalah label sebenarnya (0 atau 1 untuk setiap kelas) dari *one-hot encoding* vektor target. $\hat{y}_i$ adalah probabilitas yang diprediksi untuk kelas i. C merupakan jumlah kelas yang diprediksi.

3.2.8 Pelatihan Model

Penelitian ini akan menggunakan *optimizer* sama dengan yang digunakan Vaswani et al., yaitu *optimizer* Adam [26] yang menggabungkan momentum dengan RMSprop. Model akan dilatih menggunakan *Notebook Kaggle* dengan GPU NVIDIA Tesla P100-PCIE-16GB untuk model kecil dengan jumlah blok *encoder* 6 dan 2 GPU NVIDIA Tesla T4-15GB untuk model besar dengan jumlah blok *encoder* 12. Beberapa kedalaman *layer* dan jumlah *head attention* akan diuji performanya dengan tetap memperhatikan batasan komputasi.

Algorithm 1 Pelatihan Model

Inisialisasi: $m_0 \leftarrow 0, v_0 \leftarrow 0, t \leftarrow 0, \beta_1 \leftarrow 0.9, \beta_2 \leftarrow 0.98, \epsilon \leftarrow 10^{-9}$
Inisialisasi: epochs, batch_size
while epoch → epochs **do**
    while step → N // batch_size **do**
        $lr \leftarrow (d_{embed}^{-0.5} \cdot min(step^{-0.5},\ step \times 4000^{-1.5}))$
        $t \leftarrow t + 1$
        $g_t \leftarrow \nabla_\theta \mathcal{L}(\theta_{t-1})$ (Gradien *loss function*)
        $m_t \leftarrow \beta_1 \cdot m_{t-1} + (1 - \beta_1) \cdot g_t$ (Update momentum pertama)
        $v_t \leftarrow \beta_2 \cdot v_{t-1} + (1 - \beta_2) \cdot g_t^2$ (Update momentum kedua)
        $\hat{m}_t \leftarrow m_t / (1 - \beta_1^t)$ (Koreksi bias momentum pertama)
        $\hat{v}_t \leftarrow v_t / (1 - \beta_2^t)$ (Koreksi bias momentum kedua)
        $\theta_t \leftarrow \theta_{t-1} - lr \cdot \hat{m}_t / (\sqrt{\hat{v}_t + \epsilon})$ (Update parameter)
    end while
end while
***Return*:** Parameter θ_t

3.3 Evaluasi Model

3.3.1 *Confusion Matrix*

Confussion matrix adalah metode dalam pembelajaran mesin yang menyediakan representasi visual dari performa model dalam tugas klasifikasi [27]. Matriks ini merangkum prediksi yang benar dan salah yang dibuat oleh model, yang memungkinkan penghitungan berbagai metrik kinerja seperti akurasi, presisi, *recall*, dan *F1-score*. Matriks ini biasanya terdiri dari empat komponen: *True Positive*, *True Negarive*, *False Positive*, dan *False Negative* 3.2.

Tabel 3.2 *Confusion Matrix* untuk 3 Kelas (a), Rumus metrik evaluasi (b)

Actual Class		Predicted Class		
		C_1	C_2	C_3
Actual	C_1	TP_1	FP_{12}	FP_{13}
	C_2	FN_{21}	TP_2	FP_{23}
	C_3	FN_{31}	FN_{32}	TP_3

(a)

Metrik	Rumus
Accuracy	$\frac{TP+TN}{TP+TN+FP+FN}$
Precision	$\frac{TP}{TP+FP}$
Recall	$\frac{TP}{TP+FN}$
F1-Score	$2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision+Recall}}$

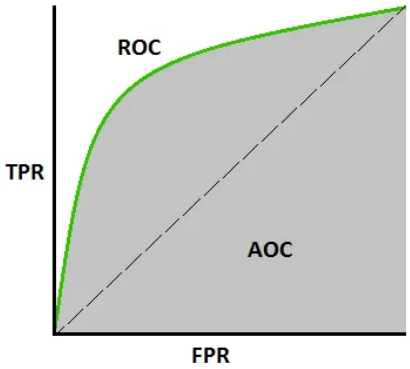
(b)

3.3.2 Kurva ROC-AUC

Kurva ROC-AUC merupakan metode untuk mengevaluasi keakuratan algoritma klasifikasi, yang memberikan wawasan tentang kinerjanya di berbagai ambang batas. Kurva ini mengukur keseimbangan antara sensitivitas (*true positive rate*) dan spesifisitas (*false positive rate*), yang memungkinkan penilaian komprehensif terhadap efektivitas model [28].

$$\text{TPR/sensitivitas} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{3.2}$$

$$\text{FPR/spesifisitas} = \frac{\text{FP}}{\text{FP} + \text{TN}} \tag{3.3}$$



Gambar 3.4 Kurva ROC-AUC

BAB IV

HASIL DAN PEMBAHASAN

4.1 Isi Bab Hasil dan Pembahasan

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

4.2 Melampirkan Tabel

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language. 4.1.

Tabel 4.1 Parameter kelulusan tugas akhir

No.	Parameter	Nilai
1.	Penulisan	A
2.	Penulisan	A

4.2.1 Subsubbab 2 2

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written

in of the original language. There is no need for special content, but the length of words should match the language.

4.3 Subab 3

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language. 4.1

$$x + 2 = 159 \tag{4.1}$$

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

Gambar 4.1 Contoh Gambar Komputer

BAB V

KESIMPULAN DAN SARAN

5.1 Kesimpulan

Berdasarkan hasil analisis dan pengujian fungsional aplikasi ini, didapat kesimpulan sebagai berikut:

1. Lorem ipsum is a pseudo-Latin text used in web design, typography, layout, and printing in place of English to emphasise design elements over content.
2. It's also called placeholder (or filler) text. It's a convenient tool for mock-ups.
3. It helps to outline the visual elements of a document or presentation, eg typography, font, or layout. Lorem ipsum is mostly a part of a Latin text by the classical author and philospher Cicero.
4. Its words and letters have been changed by addition or removal, so to deliberately render its content nonsensical; it's not genuine, correct, or comprehensible Latin anymore.

5.2 Saran

Hal-hal penting terkait pelaksanaan penelitian yang perlu diperhatikan kedepannya adalah

1. Lorem ipsum is a pseudo-Latin text used in web design, typography, layout, and printing in place of English to emphasise design elements over content.
2. It's also called placeholder (or filler) text. It's a convenient tool for mock-ups.
3. It helps to outline the visual elements of a document or presentation, eg typography, font, or layout. Lorem ipsum is mostly a part of a Latin text by the classical author and philospher Cicero.
4. Its words and letters have been changed by addition or removal, so to deliberately render its content nonsensical; it's not genuine, correct, or comprehensible Latin anymore.

DAFTAR PUSTAKA

[1] M. A. Fadhilah, "Chronic obstructive pulmonary disease", *Jurnal Medika Nusantara*, vol. 2, no. 2, hlmn. 117–125, Mei 2024. sumber: https://jurnal.stikeskesdam4dip.ac.id/index.php/Medika/article/view/1127.

[2] WHO, *Chronic obstructive pulmonary disease (COPD) — who.int*, https://www.who.int/news-room/fact-sheets/detail/chronic-obstructive-pulmonary-disease-(copd), [Accessed 27-11-2024], 2024.

[3] W. EMRO, *WHO EMRO | Chronic obstructive pulmonary disease (COPD) | Health topics — emro.who.int*, https://www.emro.who.int/health-topics/chronic-obstructive-pulmonary-disease-copd/index.html, [Accessed 27-11-2024], 2024.

[4] *Pasien PPOK RI 19 Juta di 2024, Diprediksi Terus Meningkat - Gaya Hidup — bloombergtechnoz.com*, https://www.bloombergtechnoz.com/detail-news/55485/pasien-ppok-ri-19-juta-di-2024-diprediksi-terus-meningkat, [Accessed 27-11-2024].

[5] J. Heitmann, A. Glangetas, J. Doenz, dkk., "Deepbreath—automated detection of respiratory pathology from lung auscultation in 572 pediatric outpatients across 5 countries", *NPJ digital medicine*, vol. 6, no. 1, hlmn. 104, 2023.

[6] Y. Ma, X. Xu, dan Y. Li, "Lungrn+ nl: An improved adventitious lung sound classification using non-local block resnet neural network with mixup data augmentation.", di dalam *Interspeech*, 2020, hlmn. 2902–2906.

[7] K. He, X. Zhang, S. Ren, dan J. Sun, "Deep residual learning for image recognition", di dalam *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, hlmn. 770–778.

[8] S. Gairola, F. Tom, N. Kwatra, dan M. Jain, "Respirenet: A deep neural network for accurately detecting abnormal lung sounds in limited data setting", di dalam *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, IEEE, 2021, hlmn. 527–530.

[9] B. Elizalde, S. Deshmukh, dan H. Wang, *Natural language supervision for general-purpose audio representations*, 2024. arXiv: 2309.05767 [cs.SD]. sumber: https://arxiv.org/abs/2309.05767.

[10] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lučić, dan C. Schmid, *Vivit: A video vision transformer*, 2021. arXiv: 2103.15691 [cs.CV]. sumber: https://arxiv.org/abs/2103.15691.

[11] H. Lin, X. Cheng, X. Wu, dan D. Shen, "Cat: Cross attention in vision transformer", di dalam *2022 IEEE International Conference on Multimedia and Expo (ICME)*, 2022, hlmn. 1–6.

[12] H. Xue dan F. D. Salim, "Exploring self-supervised representation ensembles for covid-19 cough classification", di dalam *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, Ser. KDD '21, ACM, Agt. 2021, hlmn. 1944–1952. sumber: http://dx.doi.org/10.1145/3447548.3467263.

[13] L. Xiao, L. Fang, Y. Yang, dan W. Tu, "Lungadapter: Efficient adapting audio spectrogram transformer for lung sound classification", di dalam *Proc. Interspeech 2024*, 2024, hlmn. 4738–4742.

[14] V. Basu dan S. Rana, "Respiratory diseases recognition through respiratory sound with the help of deep neural network", di dalam *2020 4th International Conference on Computational Intelligence and Networks (CINE)*, 2020, hlmn. 1–6.

[15] C. Barstow dan D. Forbes, "Respiratory conditions: Chronic obstructive pulmonary disease", *FP essentials*, vol. 486, hlmn. 26–32, Nov. 2019. sumber: http://europepmc.org/abstract/MED/31710455.

[16] D. Singh, M. Miravitlles, dan C. Vogelmeier, "Chronic obstructive pulmonary disease individualized therapy: Tailored approach to symptom management", *Advances in therapy*, vol. 34, hlmn. 281–299, 2017.

[17] D. M. Mannino, "Chronic obstructive pulmonary disease: Epidemiology and evaluation", *Hospital physician*, vol. 37, no. 10, hlmn. 22–40, 2001.

[18] L. D. Benton, "Childhood respiratory conditions: Asthma", *FP essentials*, vol. 513, hlmn. 11–19, Feb. 2022. sumber: http://europepmc.org/abstract/MED/35143150.

[19] M. Malarvili, T. A. Howe, S. Ramanathan, M. Alexie, dan O. P. Singh, "Chapter two - asthma: The disease and issues in monitoring the asthmatic attack", di dalam *Systems and Signal Processing of Capnography as a Diagnostic Tool for Asthma Assessment*, M. Malarvili, T. A. Howe, S. Ramanathan, M. Alexie, dan O. P. Singh, timed., Academic Press, 2023, hlmn. 25–50. sumber: https://www.sciencedirect.com/science/article/pii/B9780323857475000073.

[20] S. M. Al Sasongko, S. Tsaury, S. Ariessaputra, dan S. Ch, "Mel frequency cepstral coefficients (mfcc) method and multiple adaline neural network model for speaker identification", *JOIV: International Journal on Informatics Visualization*, vol. 7, no. 4, hlmn. 2306–2312, 2023.

[21] J. Sueur, “Mel-frequency cepstral and linear predictive coefficients”, di dalam *Sound Analysis and Synthesis with R*. Cham: Springer International Publishing, 2018, hlmn. 381–398. sumber: https://doi.org/10.1007/978-3-319-77647-7_12.

[22] A. Vaswani, N. Shazeer, N. Parmar, dkk., *Attention is all you need*, 2023. arXiv: 1706.03762 [cs.CL]. sumber: https://arxiv.org/abs/1706.03762.

[23] A. Dosovitskiy, L. Beyer, A. Kolesnikov, dkk., *An image is worth 16x16 words: Transformers for image recognition at scale*, 2021. arXiv: 2010.11929 [cs.CV]. sumber: https://arxiv.org/abs/2010.11929.

[24] H. Kaushik, *Medical sound classification challenge*, https://kaggle.com/competitions/airs-ai-in-respiratory-sounds, Kaggle, 2024.

[25] M. Avadhani, Jahnavi, A. P. Bidargaddi, dan T. S, “Multi-class urban sound classification with deep learning architectures”, di dalam *2024 5th International Conference for Emerging Technology (INCET)*, 2024, hlmn. 1–7.

[26] K. Ahn, Z. Zhang, Y. Kook, dan Y. Dai, *Understanding adam optimizer via online learning of updates: Adam is ftrl in disguise*, 2024. arXiv: 2402.01567 [cs.LG]. sumber: https://arxiv.org/abs/2402.01567.

[27] E. Manai, M. Mejri, dan J. Fattahi, “Confusion matrix explainability to improve model performance: Application to network intrusion detection”, di dalam *2024 10th International Conference on Control, Decision and Information Technologies (CoDIT)*, 2024, hlmn. 1–5.

[28] A. D. Rahajoe, Agussalim, R. Mumpuni, dkk., “Optimization of binary classification based on receiver operating characteristic area under the curve for supervised machine learning”, di dalam *2023 IEEE 9th Information Technology International Seminar (ITIS)*, 2023, hlmn. 1–6.

[29] K. S. Rao dan K. Manjunath, *Speech recognition using articulatory and excitation source features*. Springer, 2017.

[30] H. A. Feldman, N. Kaiser, dan J. A. Peacock, “Power spectrum analysis of three-dimensional redshift surveys”, *arXiv preprint astro-ph/9304022*, 1993.

[31] H. Yin, V. Hohmann, dan C. Nadeu, “Acoustic features for speech recognition based on gammatone filterbank and instantaneous frequency”, *Speech Communication*, vol. 53, no. 5, hlmn. 707–715, 2011, Perceptual and Statistical Audition. sumber: https://www.sciencedirect.com/science/article/pii/S016763931000009%2019.

[32] G. Strang, “The discrete cosine transform”, *SIAM review*, vol. 41, no. 1, hlmn. 135–147, 1999.

[33] D. Hendrycks dan K. Gimpel, *Gaussian error linear units (gelus)*, 2023. arXiv: 1606.08415 [cs.LG]. sumber: https://arxiv.org/abs/1606.08415.

LAMPIRAN

LAMPIRAN A

Mel frequency Capstral Coefficients (MFCC)

A.1 Pre-Emphasis

Pre-Emphasis merupakan salah satu praktik pra-pemrosesan umum dalam bidang pemrosesan sinyal yang digunakan untuk mengompensasi frekuensi tinggi sinyal yang ditekan selama produksi sinyal. Pra-penekanan merupakan langkah pertama selama adaptasi MFCC, yang dapat diadopsi hanya dengan menerapkan filter high-pass dengan pengaturan $[1, -0.97]$. Proses penyaringan mengubah distribusi energi di seluruh frekuensi, serta tingkat energi keseluruhan. Formula Pre-Emphasis dapat dilihat pada persamaan A.1.

$$\hat{y}[n] = x[n] - \alpha \cdot x[n-1] \tag{A.1}$$

Keterangan:

- $x[n]$ adalah sinyal input
- $\hat{y}[n]$ adalah sinyal keluaran
- α adalah faktor pre-emphasis

Keluaran $\hat{y}[n]$ akan menjadi seperti persamaan A.2

$$\{x(0), x(1) - \alpha x(0), \ldots, x(n-1) - ax(n-2)\} \tag{A.2}$$

A.2 *Framing* dan *Windowing*

Ide di balik pemisahan sinyal menjadi "frame" yang berbeda adalah memecah sinyal data mentah menjadi frame yang sinyalnya cenderung lebih stasioner. Untuk karakteristik akustik yang stabil, suara perlu diperiksa dalam jangka waktu yang cukup singkat. Oleh karena itu, pengukuran spektral jangka pendek biasanya dilakukan selama jendela 20 ms, dan setiap bingkai tumpang tindih 10 ms dengan bingkai berikutnya. Tumpang tindih bingkai sebesar 10 ms memungkinkan karakteristik temporal sinyal audio dilacak. Dengan tumpang tindih bingkai audio, representasi suara akan kira-kira terpusat pada beberapa bingkai. Pada setiap frame, jendela diterapkan untuk mempersempit sinyal ke arah batas frame. Secara umum, jendela Hanning dan Hamming [29] adalah salah satu metode yang paling banyak digunakan. Jendela ini dapat meningkatkan harmonik, menghaluskan tepi, dan mengurangi efek tepi saat melakukan DFT pada sinyal. Framing dan windowing

diterapkan dengan persamaan A.3.

$$x_n[m] = x[n] \cdot w[n] \tag{A.3}$$

Keterangan:

- $w[n]$ adalah fungsi windowing
- $x_n[m]$ adalah hasil windowing
- $x[n]$ adalah sinyal input

A.3 *Discrete Fourier Transform* (DFT)

Discrete Fourier Transform (DFT) banyak digunakan untuk menghitung spektrum daya. Spektrum daya dapat dideskripsikan sebagai distribusi dari daya pada komponen frekuensi pada sinyal [30]. Spektrum daya masing-masing frame dapat ditentukan dengan persamaan A.4.

$$X[k] = \sum_{n=0}^{N-1} x[n] \cdot e^{-j2\pi kn/N} \tag{A.4}$$

Keterangan:

- j adalah bilangan imajiner
- N adalah panjang sinyal
- $x[n]$ adalah sinyal input

dengan $x(n)$ adalah sinyal diskrit dan N adalah panjang dari sinyal.

A.4 *Mel-Frequency Filter Bank*

Mel-Frequency Filter Bank adalah bank filter yang dibangun berdasarkan persepsi nada. Filter Mel awalnya dikembangkan untuk *speech recognition* dan seperti persepsi telinga manusia terhadap ucapan, filter ini menargetkan ekstraksi representasi nonlinier dari sinyal ucapan. *Mel-Frequency Filter Bank* konvensional dibangun dari 40 filter segitiga [31]. Fungsi transfer (TF) dari masing-masing filter ke-m dapat dihitung melalui persamaan A.5,

$$H_m(k) = \begin{cases} 0 & k < f(m-1) \\ \dfrac{k - f(m-1)}{f(m) - f(m-1)} & f(m-1) \leq k < f(m) \\ 1 & k = f(m) \\ \dfrac{f(m+1) - k}{f(m+1) - f(m)} & f(m) < k \leq f(m+1) \\ 0 & k > f(m+1) \end{cases} \tag{A.5}$$

dengan $f(m)$ adalah pusat frekuensi dari filter segitiga dan $\Sigma_m^{M-1} H_m(k) = 1$. Skala Mel terhadap frekuensi respons dan sebaliknya dihitung dengan persamaan A.6 dan A.7.

$$m = 2595 \log_{10}\left(1 + \frac{f}{700}\right) \tag{A.6}$$

$$f = 700\left(10^{\frac{m}{2595}} - 1\right) \tag{A.7}$$

Keterangan:

- f adalah frekuensi
- m adalah skala mel

A.5 *Discrete Cosine Transform* (DCT)

Discrete Cosine Transform (DCT) menyatakan *finite sequence* dari titik data mengenai penjumlahan fungsi kosinus yang berosilasi pada frekuensi yang berbeda. DCT diperkenalkan oleh Nasir Ahmed pada tahun 1972. Dalam proses MFCC, DCT diterapkan pada bank filter Mel untuk memilih koefisien yang paling akurat atau untuk memisahkan hubungan dalam besaran spektral logaritma dari bank filter [32]. DCT dihitung dengan persamaan A.8

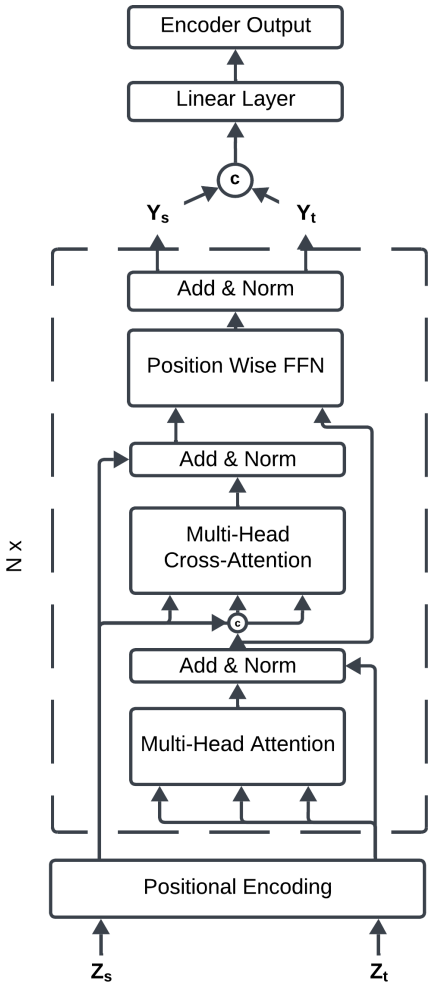
$$X_k = \sum_{n=0}^{N-1} x_n \cos\left(\frac{2\pi jnk}{N}\right), \quad k = 0, 1, \ldots, N-1 \tag{A.8}$$

Keterangan:

- x_n adalah sinyal diskrit
- N adalah panjang sinyal
- X_k adalah koefisien MFCC

LAMPIRAN B

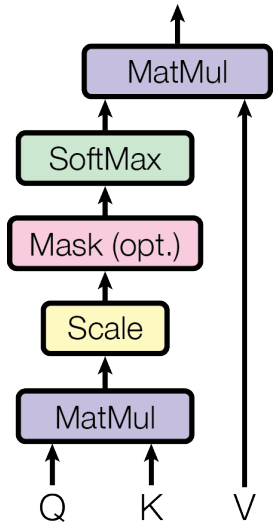
Perhitungan *Encoder Cross-Attention*



Gambar B.1 Encoder Cross Attention

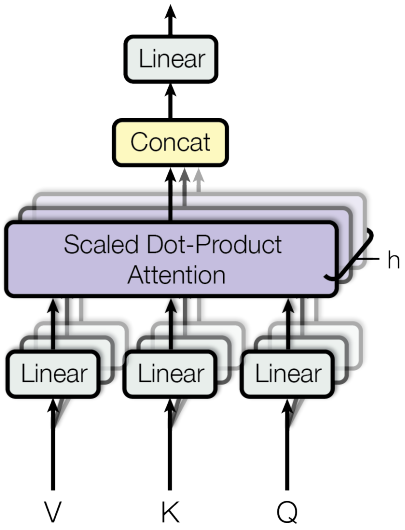
B.1 *Self-Attention* dan *Multi-Head Attention*

Proses dari *Self-Attention* atau yang memiliki nama lain *Scaled Dot-Product Attention* dapat dilihat pada Gambar B.2.



Gambar B.2 *Self-Attention* atau *Scaled Dot-Product Attention*

Proses dari *Multi-Head Attention* dapat dilihat pada Gambar B.3.



Gambar B.3 *Multi-Head Attention*

Inisiasi nilai variabel diperlukan untuk penentuan nilai awal pada pelatihan. Nilai yang digunakan adalah nilai yang sama dengan paper *"Attention Is All You Need"* dengan beberapa penyesuaian [22]. Nilai variabel inisiasi dapat dilihat pada tabel B.1.

Tabel B.1 Tabel Variabel untuk Encoder Transformator

Nama Variabel	Nilai Variabel
EMBED_DIM	64
SEQ_LENGTH	512
NUM_FEATURE	8
D_MODELS	8
NUM_HEADS	8
DROPOUT_RATE	0.2
EPOCHS	30

Matriks fitur riwayat pasien Z_t berukuran 8×1 dimasukkan ke dalam layer linear agar memiliki dimensi yang sama dengan matriks hasil ekstraksi fitur MFCC sehingga memiliki ukuran 8×64. Matriks fitur suara Z_s hasil ekstraksi fitur MFCC berukuran 512×64 dan matriks fitur riwayat pasien berukuran 8×64 ditambahkan dengan token [cls] menjadi masing-masing berukuran 513×64 dan 9×64.

Matriks fitur riwayat pasien lalu masuk ke dalam mekanisme *Multi-Head Attention* dan *Layer Norm* dengan hasil matriks berukuran 9×64. Matriks ini selanjutnya di *concat* dengan matriks fitur suara yang telah ditambah token [cls] berukuran 513×64 sehingga menghasilkan matriks berukuran 521×64.

Matriks fitur suara yang telah di *concat* dengan matriks fitur riwayat pasien yang telah melalui mekanisme *Multi-Head Attention* lalu masuk ke dalam mekanisme *Multi-Head Cross Attention*. Operasi yang digunakan pada *Multi-Head Cross Attention* sama dengan *Multi-Head Attention*, perbedaannya terletak pada input berupa gabungan antara fitur suara dan fitur riwayat pasien. Setelah melewati *Layer Norm* ukuran matriks hasil *Multi-Head Cross Attention* menjadi 513×64.

B.2 *Feed-Forward Network*

Matriks hasil *Multi-Head Attention* dan *Multi-Head Cross Attention* masing-masing masuk ke dalam mekanisme *Feed-Forward Network*. mekanisme *Feed-Forward Network* terdiri dari tiga operasi utama yaitu *Linear Transformation*, *Activation Function* dan *Layer Normalization*. Transformasi Linear diimplementasikan sebagai *fully connented layer*, atau juga dikenal sebagai *dense layer*, yang menghubungkan setiap neuron masukan ke setiap neuron keluaran. Langkah berikutnya dalam operasi *Feed-Forward Network* adalah menerapkan fungsi aktivasi. Fungsi ini merupakan fungsi non-linier yang memungkinkannya mempelajari pola yang lebih kompleks. Fungsi aktivasi yang digunakan pada

model ini adalah GELU (*Gaussian Error Linear Unit*) yang memiliki kecepatan dan konvergensi yang lebih baik karena telah menggabungkan properti dari dropout, zoneout, dan ReLUs [33]. Persamaan fungsi aktivasi GELU dapat dilihat pada persamaan B.1.

$$\text{GELU}(x) = xP(X \leq x) = x\Phi(x) = x \cdot \frac{1}{2}\left[1 + \text{erf}(x/\sqrt{2})\right] \tag{B.1}$$

Fungsi aktivasi GELU dapat diaproksimasi dengan persamaan B.2

$$0.5x(1 + \tanh[\sqrt{2/\pi}(x + 0.044715x^3)]) \tag{B.2}$$

atau persamaan B.3.

$$x\sigma(1.702x) \tag{B.3}$$

Tahap terakhir pada *Feed-Forward Network* adalah *Layer Normalization*. *Layer Normalization* adalah teknik yang menormalkan masukan di seluruh dimensi fitur (bukan dimensi batch), menstabilkan jaringan dan mempercepat pelatihan. Output dari *Feed-Forward Network* berupa kedua matriks dengan ukuran 513×64 dan 9×64.

B.3 Output Encoder Y_s dan Y_t

Proses dari keseluruhan di atas merupakan proses dalam satu blok encoder, proses ini diulang sebanyak N_BLOCK. Didapatkan vektor Y_s dan Y_t dengan ukuran 1×64 yang kemudian di *concat* secara horizontal/kolom sehingga didapatkan vektor berukuran 1×128. Vektor ini kemudian masuk ke dalam fungsi aktivasi GELU dan layer Linear sehingga didapatkan vektor berukuran 1×64. Fungsi aktivasi Softmax digunakan untuk mengakomodir klasifikasi multi-kelas dengan menghasilkan probabilitas setiap kelas dengan jumlah semua probabilitas berjumlah tepat 1. Fungsi aktivasi Softmax dapat dilihat pada persamaan B.4.

$$\sigma(\mathbf{z})_i = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}}, \quad \text{for } i = 1, \ldots, K \tag{B.4}$$

Keterangan:

- σ adalah softmax
- $\mathbf{z}$ adalah input vektor
- K adalah jumlah kelas
- e^{z_i} fungsi eksponensial untuk vektor input

- e^{z_j} fungsi eksponensial untuk vektor output

Fungsi B.4 menghasilkan vektor 1×3 yang berisi nilai probabilitas dari ketiga kelas. Nilai probabilitas terbesar menunjukkan prediksi kelas dari data.

LAMPIRAN C

Optimizer Adam

C.1 Optimizer Adam pada Transformers

Adam menggabungkan kelebihan dari dua algoritma optimisasi yaitu momentum dan RMSProp dengan mengestimasi rata-rata dan varians dari gradien. Dalam model Transformers, parameter θ mencakup bobot dan bias untuk:

- Lapisan *self-attention* dan *multi-head attention* (seperti bobot proyeksi query, key dan value).
- *Feed-Forward Network*.
- Parameter *Layer Normalization*.

Setiap parameter diperbarui menggunakan persamaan di bawah.

C.1.1 Persamaan Oprimizer Adam

1. Menghitung gradien dari loss L dengan parameter θ dengan persamaan C.1:

$$g_t = \nabla_\theta L(\theta_t) \tag{C.1}$$

2. Perbarui estimasi momentum bias dengan persamaan C.2dan C.3:

$$m_t = \beta_1 m_{t-1} + (1 - \beta_1) g_t \tag{C.2}$$

$$v_t = \beta_2 v_{t-1} + (1 - \beta_2) g_t^2 \tag{C.3}$$

3. Terapkan koreksi bias pada momentum dengan persamaan C.4:

$$\hat{m}_t = \frac{m_t}{1 - \beta_1^t}, \quad \hat{v}_t = \frac{v_t}{1 - \beta_2^t} \tag{C.4}$$

4. Perbarui parameter dengan persamaan C.5:

$$\theta_{t+1} = \theta_t - \eta \frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \epsilon} \tag{C.5}$$

Keterangan:

- g_t adalah fungsi gradien
- $L(\theta_t)$ adalah fungsi Loss

- m_t adalah momentum
- v_t adalah velocity
- η merupakan *learning rate*.
- β_1 dan β_2 adalah *hyperparameter* yang mengendalikan tingkat peluruhan estimasi momen.
- ϵ adalah konstanta kecil untuk mencegah pembagian dengan nol.

C.1.2 Contoh Perhitungan

Asumsikan :

- Learning rate $\eta = 0.001$,
- $\beta_1 = 0.9, \beta_2 = 0.999$,
- $\epsilon = 10^{-8}$,
- Gradien pada *timestep* t: $g_t = 0.02$,
- Estimasi momentum sebelumnya: $m_{t-1} = 0.01$, $v_{t-1} = 0.0001$,
- $t = 10$.

Langkah-langkah perhitungan:

1. Hitung estimasi momen bias:

$$m_t = \beta_1 m_{t-1} + (1 - \beta_1) g_t = 0.9 \cdot 0.01 + (1 - 0.9) \cdot 0.02 = 0.011$$

$$v_t = \beta_2 v_{t-1} + (1 - \beta_2) g_t^2 = 0.999 \cdot 0.0001 + (1 - 0.999) \cdot (0.02)^2 = 0.0001004$$

2. Terapkan koreksi bias:

$$\hat{m}_t = \frac{m_t}{1 - \beta_1^t} = \frac{0.011}{1 - 0.9^{10}} \approx 0.0111$$

$$\hat{v}_t = \frac{v_t}{1 - \beta_2^t} = \frac{0.0001004}{1 - 0.999^{10}} \approx 0.000102$$

3. Hitung pembaruan parameter:

$$\Delta\theta = \eta \frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \epsilon}$$

Subsitusikan nilai:

$$\Delta\theta = 0.001 \cdot \frac{0.0111}{\sqrt{0.000102} + 10^{-8}} \approx 0.00109$$

4. Perbarui parameter:

$$\theta_{t+1} = \theta_t - \Delta\theta$$

LAMPIRAN D

Variasi Model dan *Hyperparameter Tuning*

Tabel D.1 Variasi Model dan Hyperparameter Transformer

Hyperparameter	**Small Model**	**Big Model**
Jumlah Layer/Blok Encoder (*N_BLOCK*)	6	12
Dimensi Embedding (d_{model})	64	128
Dimensi Feedforward (d_{ff})	256	512
Jumlah Head Attention	8	16
Dropout	0.2	0.3
Batch Size	32	64
Learning Rate	1e-4	5e-5