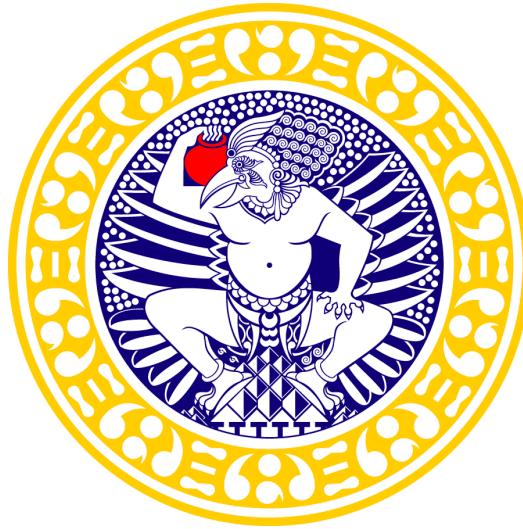


# **LAPORAN *FINAL PROJECT***

## **Analisis Suara Paru-Paru untuk Sistem Deteksi Penyakit Paru-Paru Menggunakan *Machine Learning***



Disusun oleh:  
Kelompok 6

- |                         |           |
|-------------------------|-----------|
| 1. Arkan Syafiq At'taqy | 164221062 |
| 2. Kyla Belva Queena    | 164221015 |
| 3. Ramadhan Eko Saputra | 164221088 |
| 4. Rizal Dwi Prasetyo   | 164221059 |

**SD-A1**

**MATA KULIAH DATA MINING II  
PROGRAM STUDI TEKNOLOGI SAINS DATA  
FAKULTAS TEKNOLOGI MAJU DAN MULTIDISIPLIN  
UNIVERSITAS AIRLANGGA**

**2024**

## DAFTAR ISI

DAFTAR ISI.....	2
BAB I.....	4
PENDAHULUAN.....	4
1.1 Latar Belakang.....	4
1.2 Rumusan Masalah.....	4
1.3 Tujuan Penelitian.....	5
1.4 Manfaat Penelitian.....	5
1.5 Batasan Penelitian.....	5
BAB II.....	6
TINJAUAN PUSTAKA.....	6
2.1 Klasifikasi.....	6
2.2 Penyakit Paru Obstruktif Kronik (PPOK).....	6
2.3 Mendeley Data sebagai Sumber Data.....	6
2.4 Exploratory Data Analysis (EDA).....	7
2.5 Data Preprocessing.....	7
2.6 Ekstraksi Fitur Audio.....	7
2.6.1 Mel-frequency Cepstral Coefficients (MFCC).....	8
2.6.2 Spectral Roll-Off.....	8
2.6.3 Zero-Crossing Rate.....	8
2.7 Model Klasifikasi Machine Learning.....	9
2.7.1 Random Forest.....	9
2.7.2 Light Gradient Boosting Machine (LGBM).....	9
2.7.3 Support Vector Machine (SVM).....	10
2.7.4 Voting Classifier.....	10
BAB III.....	11
METODOLOGI.....	11
3.1 Sumber Data.....	11
3.1.1 Pengambilan Data.....	11
3.1.2 Variabel Penelitian.....	11
3.2 Diagram Alur Penelitian.....	11
3.3 Exploratory Data Analysis (EDA).....	12
3.3.1 Pie Chart.....	12
3.3.2 Waveform.....	12
3.3.3 Spectrum.....	12
3.3.4 Spectrogram.....	12
3.3.5 MFCC.....	12
3.4 Preprocessing Data.....	12
3.4.1 Feature Extraction.....	12
3.4.2 Load Data.....	13
3.4.3 Encoding.....	13
3.4.4 Train-Test Split.....	13
3.5 Metode Klasifikasi.....	13

3.6 Evaluasi Model.....	14
3.6.1 Classification Report.....	14
3.6.2 Confusion Matrix.....	14
BAB IV.....	16
HASIL DAN ANALISIS.....	16
4.1 Eksplorasi Data Analysis (EDA).....	16
4.1.1 Distribusi Data.....	16
4.1.2 Waveform.....	16
4.1.3 Spectrum.....	17
4.1.4 Spectrogram.....	17
4.1.5 MFCCs.....	18
4.2 Modelling.....	18
4.2.1 Random Forest.....	18
4.2.2 Light Gradient Boost Machine (LGBM).....	19
4.2.3 Support Vector Machine (SVM).....	19
4.2.4 Voting Classifier.....	19
4.2.5 Pemilihan Model Terbaik.....	20
4.3 Sistem Deteksi Penyakit Paru-Paru.....	20
BAB V.....	21
KESIMPULAN DAN SARAN.....	21
5.1 Kesimpulan.....	21
5.2 Saran.....	21
DAFTAR PUSTAKA.....	22

## DAFTAR TABEL

Tabel 1 .....	11
Tabel 2 .....	15
Tabel 3 .....	16
Tabel 4 .....	16
Tabel 5 .....	17
Tabel 6 .....	17
Tabel 7 .....	19
Tabel 8 .....	19
Tabel 9 .....	19
Tabel 10 .....	19
Tabel 11 .....	19

## DAFTAR GAMBAR

Gambar 1 .....	11
Gambar 2 .....	15

# BAB I

## PENDAHULUAN

### 1.1 Latar Belakang

Penyakit paru-paru, khususnya Penyakit Paru Obstruktif Kronik (PPOK), menjadi salah satu penyebab utama kematian di dunia. Berdasarkan data WHO, PPOK menempati urutan keempat sebagai penyebab kematian terbanyak global, dengan lebih dari 3,5 juta orang meninggal akibat penyakit ini pada tahun 2021. Salah satu tantangan utama dalam mengurangi angka kematian ini adalah keterlambatan dalam diagnosis, karena banyak pasien baru mengunjungi dokter pada tahap lanjut penyakit. Penggunaan teknologi dalam mendiagnosis penyakit paru-paru, terutama di daerah dengan sumber daya terbatas menjadi sangat penting.

Salah satu metode yang dapat diterapkan untuk mempercepat diagnosis adalah menggunakan teknologi berbasis machine learning untuk menganalisis suara paru-paru. Dalam metode tradisional, auskultasi dengan stetoskop digunakan untuk mendengarkan suara pernapasan pasien. Namun, sistem ini memiliki keterbatasan, terutama di daerah terpencil yang kekurangan tenaga medis. Solusi berbasis machine learning, yang menganalisis suara paru-paru untuk mendeteksi pola yang menunjukkan adanya gangguan pernapasan, dapat menawarkan diagnosa cepat dan murah. Dengan menggunakan dataset seperti Mendeley Data yang menyediakan suara paru-paru, sistem ini dapat dilatih untuk mendeteksi gejala penyakit paru-paru secara otomatis.

Implementasi sistem deteksi berbasis suara paru-paru dengan machine learning memiliki potensi besar untuk meningkatkan diagnosa dini penyakit paru-paru, terutama di wilayah dengan keterbatasan sumber daya medis. Dengan menganalisis suara pernapasan menggunakan fitur seperti *Mel-Frequency Cepstral Coefficients* (MFCC), *Spectral Roll-Off*, dan *Zero-Crossing Rate*, sistem ini dapat memberikan hasil yang lebih cepat dan lebih murah dibandingkan metode tradisional. Harapannya, sistem ini dapat menjadi bagian dari transformasi digital di sektor kesehatan yang memungkinkan deteksi penyakit paru-paru lebih awal, sehingga pasien dapat segera mendapatkan penanganan medis yang lebih lanjut.

### 1.2 Rumusan Masalah

Dari latar belakang yang dijelaskan diatas, maka penulis dapat memberikan rumusan masalah sebagai berikut:

1. Bagaimana cara mengolah data suara paru-paru dari Mendeley Data untuk mendukung proses klasifikasi penyakit paru-paru?
2. Apa algoritma *machine learning* yang paling efektif dalam mengklasifikasikan suara paru-paru sebagai normal atau abnormal?
3. Sejauh mana akurasi model yang dihasilkan dalam mengklasifikasikan suara

paru-paru ke dalam kategori normal dan abnormal?

### **1.3 Tujuan Penelitian**

Dari latar belakang dan rumusan masalah yang telah dijelaskan diatas, maka penulis dapat memberikan tujuan sebagai berikut:

1. Mengolah dan menganalisis data suara paru-paru dari Mendeley Data untuk mendeteksi adanya indikasi penyakit paru-paru secara otomatis.
2. Mengembangkan model klasifikasi berbasis machine learning yang mampu mengidentifikasi suara paru-paru sebagai normal atau abnormal.
3. Mengevaluasi performa model dalam mendeteksi kelainan pada suara paru-paru dengan menggunakan metrik evaluasi yang relevan.

### **1.4 Manfaat Penelitian**

Dari latar belakang, rumusan masalah, dan tujuan yang telah dijelaskan diatas, maka manfaat yang dihasilkan antara lain:

1. Meningkatkan kemampuan dalam mengembangkan sistem berbasis *machine learning* untuk aplikasi medis, khususnya dalam mendiagnosis penyakit paru-paru menggunakan analisis suara.
2. Mempercepat dan mempermudah proses diagnosis penyakit paru-paru, terutama di daerah yang sulit mengakses fasilitas medis, dengan menggunakan teknologi analisis suara secara real-time.
3. Mendorong inovasi dalam pengolahan data suara untuk bidang kesehatan, serta meningkatkan kesadaran akan pentingnya teknologi dalam meningkatkan akurasi dan kecepatan pengambilan keputusan medis.

### **1.5 Batasan Penelitian**

Dari latar belakang, rumusan masalah, dan tujuan yang telah dijelaskan diatas, adapun batasan penelitian ini antara lain:

1. Penelitian ini hanya berfokus pada klasifikasi suara pernapasan manusia (normal dan abnormal) berdasarkan data audio yang diperoleh dari sumber yang relevan.
2. Klasifikasi dalam penelitian ini terbatas pada dua kategori suara pernapasan, yaitu normal dan abnormal, dengan fokus utama pada pengidentifikasian gangguan pernapasan yang dapat menunjukkan kondisi medis tertentu.
3. Penelitian ini hanya akan menggunakan data suara pernapasan yang dikumpulkan dalam periode waktu tertentu, yang relevan dengan konteks analisis kesehatan pernapasan.

## **BAB II**

### **TINJAUAN PUSTAKA**

#### **2.1 Klasifikasi**

Klasifikasi adalah pengelompokkan data atau objek baru ke dalam kelas atau label berdasarkan atribut-atribut tertentu. Teknik dari klasifikasi adalah dengan melihat variabel dari kelompok data yang sudah ada [1]. Klasifikasi bertujuan untuk memprediksi kelas dari suatu objek yang tidak diketahui sebelumnya. Klasifikasi terdiri dari tiga tahap, yaitu pembangunan model, penerapan model, dan evaluasi. Pembangunan model adalah membangun model menggunakan data latih yang telah memiliki atribut dan kelas. Kemudian, data-data tersebut diterapkan untuk menentukan kelas dari data atau objek yang baru. Setelah itu, data dievaluasi untuk melihat tingkat akurasi dari pembangunan dan penerapan model terhadap data baru [2]. Proses klasifikasi terdiri dari dua fase, yaitu fase training dan fase testing. Fase training adalah fase di mana data digunakan untuk membangun sebuah model sedangkan fase testing adalah pengujian model yang telah dibuat dengan data lainnya untuk mengetahui akurasi dari model tersebut [3].

#### **2.2 Penyakit Paru Obstruktif Kronik (PPOK)**

Penyakit Paru Obstruktif Kronik (PPOK) atau Chronic Obstructive Pulmonary Disease (COPD) adalah suatu penyumbatan menetap pada saluran pernapasan yang disebabkan oleh emfisema dan bronkitis kronis. Menurut American College of Chest Physicians American Society, PPOK didefinisikan sebagai kelompok penyakit paru yang ditandai dengan perlambatan aliran udara yang bersifat menetap [4]. PPOK adalah penyakit yang membentuk satu kesatuan dengan diagnosa medisnya adalah bronkitis, emfisema paru-paru, dan asma bronchial [5].

#### **2.3 Mendeley Data sebagai Sumber Data**

Mendeley Data adalah repositori data penelitian yang dirancang untuk menyimpan, mengelola, dan berbagi dataset penelitian dengan cara yang mudah diakses dan efisien. Platform ini menyediakan layanan unggahan dataset dalam berbagai format file hingga ukuran 10 GB per dataset, serta memastikan validitas dan keterlacakan data melalui pemberian DOI otomatis. Dataset dapat dipublikasikan atau dibagikan secara privat kepada kolaborator, mendukung kolaborasi penelitian yang lebih efektif. Sebagai

platform berbasis cloud, Mendeley Data juga bekerja sama dengan *Data Archiving and Networked Services* (DANS) untuk memastikan penyimpanan data jangka panjang yang andal. Keunggulan lainnya termasuk antarmuka pengguna yang sederhana dan aksesibilitas tinggi, meskipun fitur pencarian dan metrik kutipan masih memerlukan pengembangan lebih lanjut [6].

## **2.4 Exploratory Data Analysis (EDA)**

*Exploratory Data Analysis* (EDA) adalah proses menganalisis dan meringkas data untuk memahami struktur dan hubungan yang mendasari data. Dalam analisis sinyal, EDA melibatkan visualisasi dan evaluasi fitur seperti *waveform*, *spectrum*, *spectrogram*, dan *Mel-Frequency Cepstral Coefficients* (MFCC). Analisis *waveform* memungkinkan visualisasi pola amplitudo sinyal terhadap waktu, *spectrum* memberikan informasi tentang distribusi frekuensi, *spectrogram* menggambarkan perubahan frekuensi terhadap waktu, sementara MFCC membantu mengekstraksi fitur penting dari sinyal untuk analisis lebih lanjut.

## **2.5 Data Preprocessing**

*Data preprocessing* adalah tahap penting dalam persiapan analisis audio, yang melibatkan beberapa langkah utama. Proses ini mencakup konversi format audio, normalisasi volume, penghapusan *noise*, dan segmentasi untuk membagi data menjadi segmen-segmen kecil [7]. Ekstraksi fitur, seperti *Mel-Frequency Cepstral Coefficients* (MFCC), digunakan untuk mengambil representasi penting dari data audio. Selain itu, encoding mengonversi label menjadi format numerik, sementara train-test split memisahkan data menjadi set pelatihan dan pengujian untuk evaluasi model yang adil. Langkah-langkah ini bertujuan mengoptimalkan performa model dalam mengenali pola dan menghasilkan prediksi yang lebih akurat.

## **2.6 Ekstraksi Fitur Audio**

Ekstraksi fitur adalah proses penting dalam pengenalan suara untuk mengidentifikasi karakteristik unik dari sinyal audio. Metode yang umum digunakan meliputi *Mel-Frequency Cepstral Coefficients* (MFCC), *spectral roll-off*, dan *zero crossing rate* (ZCR). MFCC menangkap informasi frekuensi relevan yang menyerupai persepsi pendengaran manusia, sementara *spectral roll-off* memberikan gambaran distribusi energi spektrum, dan ZCR mengukur perubahan tanda amplitudo sinyal.

### 2.6.1 Mel-frequency Cepstral Coefficients (MFCC)

Sistem pendengaran manusia diasumsikan memproses sinyal ucapan secara non-linear. Telah diketahui bahwa komponen frekuensi rendah pada sinyal ucapan mengandung lebih banyak informasi dibandingkan komponen frekuensi tinggi. Oleh karena itu, MFCC dirancang untuk memberikan penekanan lebih pada komponen frekuensi rendah daripada yang tinggi. *Mel-frequency Cepstral Coefficients* (MFCC) adalah representasi dari spektrum daya jangka pendek dari sebuah frame ucapan menggunakan transformasi kosinus linier dari log spektrum daya pada skala frekuensi Mel yang bersifat non-linear [8]. Konversi dari frekuensi normal (f) ke frekuensi Mel (m) diberikan oleh persamaan berikut:

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

### 2.6.2 Spectral Roll-Off

Fitur *spectral roll-off* didefinisikan sebagai frekuensi dengan persentase tertentu (biasanya sekitar 90%) dari distribusi besarnya spektrum dikonsentrasikan. Oleh karena itu, jika koefisien DFT ke-m sesuai dengan *spectral roll-off* dari frame ke-i, maka memenuhi persamaan berikut:

$$\sum_{k=1}^m X_i(k) = C \sum_{k=1}^{Wf_L} X_i(k) \quad [9]$$

### 2.6.3 Zero-Crossing Rate

*Zero-Crossing Rate* (ZCR) dari frame audio adalah laju perubahan tanda sinyal dalam satu frame. Dengan kata lain, ZCR dapat diartikan sebagai berapa kali sinyal mengubah nilai, dari positif ke negatif dan sebaliknya dibagi dengan panjang *frame*. ZCR didefinisikan berdasarkan persamaan berikut:

$$Z(i) = \frac{1}{2W_L} \sum_{n=1}^{W_L} \left| \text{sgn}[x_i(n)] - \text{sgn}[x_i(n-1)] \right| \quad [10]$$

## 2.7 Model Klasifikasi *Machine Learning*

*Machine Learning*, sebagai aplikasi dari *Artificial Intelligence* (AI), memungkinkan sistem untuk belajar secara mandiri dari data training tanpa pemrograman berulang. Ini



digunakan untuk menghasilkan prediksi atau keputusan berdasarkan pengalaman masa lalu [11]. Machine Learning dapat diterapkan di berbagai bidang, termasuk medis, untuk memprediksi penyakit [12]. Dalam analisis ini, model klasifikasi yang digunakan mencakup *Random Forest*, *Light Gradient Boosting Machine (LGBM)*, dan *Support Vector Machine (SVM)*, dengan *Voting Classifier* yang menggabungkan hasil dari beberapa model untuk meningkatkan akurasi dan keandalan prediksi.

### **2.7.1 *Random Forest***

*Random Forest* adalah teknik statistik nonparametrik yang efektif untuk menganalisis sensitivitas global pada model kompleks. *Random Forest* merupakan metode meta-modeling yang menganalisis hubungan antara input dan output dengan menggunakan regresi statistik [13]. Metode ini menghasilkan estimasi pentingnya variabel melalui indeks variabel berbasis permutasi. Metode ini memiliki keunggulan dalam menangani data berdimensi tinggi, korelasi antar input, dan interaksi antar variabel. Cara kerja *Random Forest* dimulai dengan membangun sejumlah pohon keputusan (*decision trees*) dari sampel bootstrap data pelatihan. Setiap pohon dibentuk melalui pemisahan data secara berulang berdasarkan aturan tertentu, seperti nilai variabel input, untuk meminimalkan heterogenitas dalam simpul data.

### **2.7.2 *Light Gradient Boosting Machine (LGBM)***

*Light Gradient Boosting Machine (LGBM)* adalah algoritma machine learning berbasis pohon yang efisien untuk klasifikasi dan regresi, terutama pada dataset besar. Menggabungkan *Gradient Boosting Decision Tree (GBDT)* dan *Gradient-Based One-Sided Sampling (GOSS)*, LGBM mengoptimalkan kecepatan dan akurasi tanpa mengorbankan efisiensi memori [14]. Algoritma ini membangun pohon keputusan secara vertikal, dimulai dari daun dengan kehilangan paling signifikan, berbeda dengan pendekatan horizontal pada algoritma lain. LGBM telah terbukti efektif dalam mendeteksi penipuan transaksi, dengan akurasi hingga 99,03%, terutama ketika menggunakan teknik seperti SMOTE untuk mengatasi ketidakseimbangan kelas dan *hyperparameter tuning* untuk mengurangi *overfitting*. Meskipun sangat cepat dan akurat pada data besar, LGBM cenderung *overfitting* pada dataset kecil, menjadikannya lebih cocok untuk dataset berskala besar.

### 2.7.3 *Support Vector Machine (SVM)*

Algoritma *Support Vector Machine (SVM)* adalah algoritma yang bertujuan untuk menemukan *hyperplane* maksimal. *Hyperplane* merupakan suatu fungsi yang dapat memisahkan antara dua kelas. Pada prosesnya, SVM akan memaksimalkan margin atau jarak antara pola pelatihan dan batas keputusan [15]. Terdapat beberapa keunggulan dari algoritma ini antara lain memiliki performa yang bagus baik digunakan dengan jumlah data kecil maupun besar, memiliki performa yang bagus pada data yang memiliki atribut yang banyak dan mudah diimplementasikan [16]. Pada awalnya, algoritma ini hanya bisa melakukan klasifikasi biner, namun saat ini telah dikembangkan sehingga mampu digunakan untuk mengklasifikasikan beberapa kelas sekaligus. Selain digunakan untuk klasifikasi, SVM juga dapat digunakan untuk regresi dan pencarian *outlier*.

### 2.7.4 *Voting Classifier*

*Voting Classifier* adalah metode *ensemble* dalam *Machine Learning* yang menggabungkan prediksi dari beberapa model untuk meningkatkan akurasi dan stabilitas. Metode ini terdiri dari dua pendekatan utama: *hard voting* dan *soft voting*. Pada *hard voting*, prediksi akhir ditentukan berdasarkan suara mayoritas dari model, sedangkan *soft voting* mempertimbangkan probabilitas prediksi setiap model dan memilih kelas dengan probabilitas tertinggi. *Soft voting* biasanya memberikan performa lebih baik, terutama ketika model-model yang digunakan menghasilkan probabilitas yang akurat [17].

## BAB III

### METODOLOGI

#### 3.1 Sumber Data

##### 3.1.1 Pengambilan Data

Pada penelitian ini, kami melakukan analisis Klasifikasi Suara Paru-Paru untuk sistem deteksi penyakit paru-paru. Dataset yang digunakan dalam penelitian ini terdiri dari 222 rekaman suara, yang diperoleh dari platform Mendeley Data dan dapat diakses melalui tautan <https://data.mendeley.com/datasets/fr7zvy8j5s/1>. Dataset ini terbagi menjadi dua kelas, yaitu suara normal dan abnormal, dengan rincian sebagai berikut:

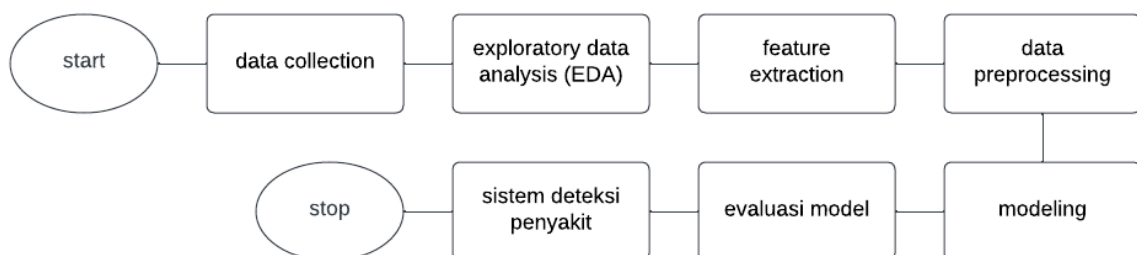
Tabel 1. Jumlah Data Masing-Masing Kelas	
Kelas	Jumlah
abnormal	197 files wav
normal	25 files wav

##### 3.1.2 Variabel Penelitian

Dataset ini memiliki dua variabel untuk proses klasifikasi suara:

1. Abnormal (88.74%) → Berisi rekaman suara dari pasien dengan kondisi penyakit paru-paru seperti COPD, *Pleural*, *Rhonching*, dan *Wheezing*.
2. Normal (11.26%) → Berisi rekaman suara dari subjek sehat tanpa kelainan paru-paru.

#### 3.2 Diagram Alur Penelitian



Gambar 1. Diagram Alur Penelitian

#### 3.3 Exploratory Data Analysis (EDA)

##### 3.3.1 Pie Chart

Pie chart merupakan visualisasi yang dipilih untuk menggambarkan distribusi masing-masing kedua kelas, yaitu normal dan abnormal.

### 3.3.2 Waveform

Menggunakan visualisasi *waveform* untuk melihat gelombang suara dengan tujuan melihat pola amplitudo sinyal terhadap waktu.

### 3.3.3 Spectrum

Melakukan analisis spektrum menggunakan FFT (*Fast Fourier Transform*) untuk memvisualisasikan komponen frekuensi dari sinyal suara.

### 3.3.4 Spectrogram

Menggunakan *spectrogram* untuk menunjukkan distribusi energi frekuensi terhadap waktu dalam sinyal suara.

### 3.3.5 MFCC

Menampilkan *Mel-Frequency Cepstral Coefficients* (MFCC) untuk mendapatkan fitur yang lebih representatif dari data suara untuk keperluan klasifikasi.

## 3.4 Preprocessing Data

### 3.4.1 Feature Extraction

Pada tahap ini, teknik ekstraksi fitur digunakan untuk mendapatkan informasi penting dari data suara paru-paru yang diperlukan untuk proses klasifikasi. *Mel-Frequency Cepstral Coefficients* (MFCC) digunakan untuk merepresentasikan frekuensi suara yang lebih sesuai untuk analisis. Pustaka Librosa digunakan untuk mengekstrak MFCC, yang menghitung koefisien spektral sinyal suara secara otomatis. Selain MFCC, dua fitur tambahan juga diekstraksi, yaitu *Spectral Roll-Off* dan *Zero-Crossing Rate* (ZCR). *Spectral Roll-Off* mengukur titik di mana energi spektral dari sinyal mulai berkurang, memberikan gambaran tentang bentuk spektrum suara. Sementara itu, ZCR menghitung jumlah perubahan tanda dalam sinyal suara, yang dapat memberikan informasi tentang tekstur atau sifat dari suara tersebut. Hasil ekstraksi dari ketiga fitur ini (MFCC, *Spectral Roll-Off*, dan ZCR) menghasilkan matriks fitur yang menunjukkan karakteristik unik dari setiap data suara dan digunakan dalam proses klasifikasi.

### 3.4.2 Load Data

Mengimpor data suara paru-paru dan dimuat ke dalam lingkungan analisis. Pada bagian ini, semua file suara dibaca dan label "normal" atau "abnormal" dimasukkan ke

dalam dataframe. Pada proses ini juga dilakukan validasi data untuk memastikan semua file telah dimuat dengan benar.

### 3.4.3 *Encoding*

Setelah data dimuat, *encoding* dilakukan untuk mengubah label kategori ("Normal" dan "Abnormal") menjadi bentuk numerik agar algoritma *machine learning* dapat memprosesnya. Pada tahap ini, metode label *encoding* langsung mengubah label menjadi angka yang berbeda, 0 untuk "Abnormal" dan 1 untuk "Normal".

### 3.4.4 *Train-Test Split*

Pada tahap ini, data dibagi menjadi dua bagian utama, yaitu data pelatihan (*training data*) dan data pengujian (*testing data*). Data dibagi menjadi 80% untuk pelatihan dan 20% untuk pengujian. Hal ini dilakukan untuk memastikan bahwa model yang akan dibuat dapat diuji pada data yang belum pernah dilihat sebelumnya untuk memungkinkan evaluasi performa model menjadi lebih objektif.

## 3.5 Metode Klasifikasi

Penelitian ini menggunakan empat algoritma pembelajaran mesin, yaitu *Random Forest*, *Light Gradient Boosting Machine* (LGBM), *Support Vector Machine* (SVM), dan *Voting Classifier*, untuk klasifikasi data audio dengan fitur *Mel-Frequency Cepstral Coefficients* (MFCC), *Spectral Roll-Off*, dan *Zero-Crossing Rate* (ZCR). *Random Forest* dipilih karena kemampuannya menangani data dengan banyak fitur dan kompleksitas melalui ensemble berbasis pohon keputusan, sementara LGBM lebih efisien dalam menangani dataset besar dan kompleks dengan menggunakan boosting iteratif. SVM digunakan karena kemampuannya dalam menangani klasifikasi dengan margin yang jelas, terutama pada data dengan batasan kelas yang tidak linier.

*Voting Classifier* diterapkan untuk menggabungkan prediksi dari semua model tersebut, memanfaatkan kekuatan masing-masing algoritma untuk meningkatkan akurasi dan stabilitas hasil klasifikasi. Proses evaluasi dilakukan dengan menggunakan *Classification Report* dan *Confusion Matrix* untuk menilai kemampuan model dalam memprediksi kategori "Normal" dan "Abnormal". Analisis dilakukan untuk membandingkan kinerja keempat model pada berbagai fitur dan kombinasi fitur, memberikan gambaran tentang model dan fitur yang paling efektif untuk klasifikasi data.

## 3.6 Evaluasi Model

### 3.6.1 Classification Report

Menyediakan metrik evaluasi seperti *precision*, *recall*, *f1-score*, dan *accuracy* untuk masing-masing kelas (Normal dan Abnormal).

a. *Accuracy*

*Accuracy* mengukur persentase prediksi yang benar (baik positif maupun negatif) dari seluruh dataset.

b. *Precision*

*Precision* mengukur seberapa tepat model dalam memprediksi kelas positif, yaitu persentase prediksi positif yang benar dari semua prediksi positif.

c. *Recall*

*Recall* mengukur seberapa baik model mendeteksi semua kasus positif, yaitu persentase true positives dari semua data yang sebenarnya positif.

d. *F1-Score*

*F1-Score* adalah rata-rata tertimbang antara *precision* dan *recall*, berguna ketika ada ketidakseimbangan antara false positives dan false negatives.

### 3.6.2 Confusion Matrix

*Confusion Matrix* digunakan untuk menganalisis kesalahan klasifikasi dan membandingkan prediksi model dengan nilai sebenarnya (*ground truth*).

a. *True Positive*

Kasus dimana model memprediksi suatu sampel sebagai positif dan prediksi tersebut benar.

b. *True Negative*

Kasus dimana model memprediksi suatu sampel sebagai negatif dan prediksi tersebut benar.

c. *False Positive*

Kasus dimana model memprediksi suatu sampel sebagai positif, tetapi prediksi tersebut salah.

d. *False Negative*

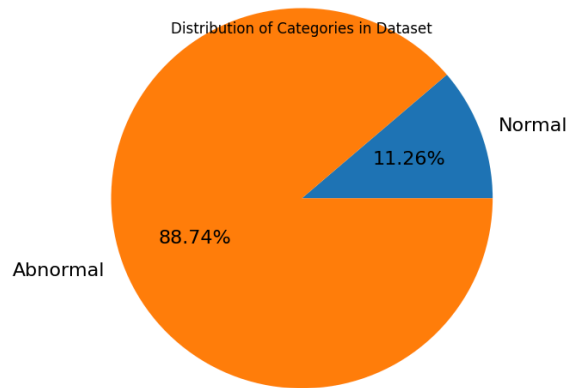
Kasus dimana model memprediksi suatu sampel sebagai negatif, tetapi prediksi tersebut salah.

## BAB IV

### HASIL DAN ANALISIS

#### 4.1 Eksplorasi Data Analysis (EDA)

##### 4.1.1 Distribusi Data

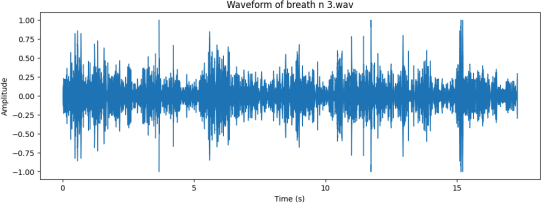
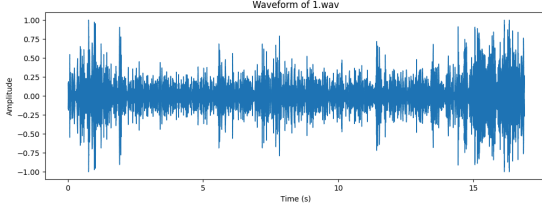


Gambar 2. Distribusi Data

Pie chart tersebut menunjukkan distribusi data dari dataset yang digunakan, di mana 88,74% data termasuk dalam kategori Abnormal, sedangkan hanya 11,26% data termasuk dalam kategori Normal.

##### 4.1.2 Waveform

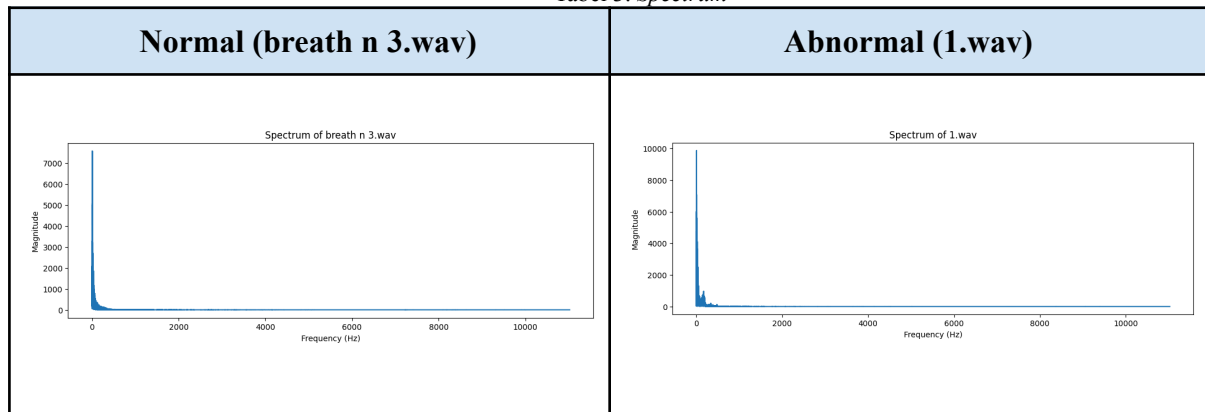
Tabel 2. Waveform

Normal (breath n 3.wav)	Abnormal (1.wav)
	

Waveform untuk "normal" menunjukkan pola suara yang stabil dengan fluktuasi amplitudo alami dalam rentang -1 hingga 1, tanpa lonjakan ekstrem, yang menandakan tidak ada gangguan signifikan. Sebaliknya, waveform "abnormal" menunjukkan ketidakteraturan, adanya *noise*, dan pola yang kurang konsisten, mengindikasikan kemungkinan gangguan pada saluran pernapasan.

### 4.1.3 Spectrum

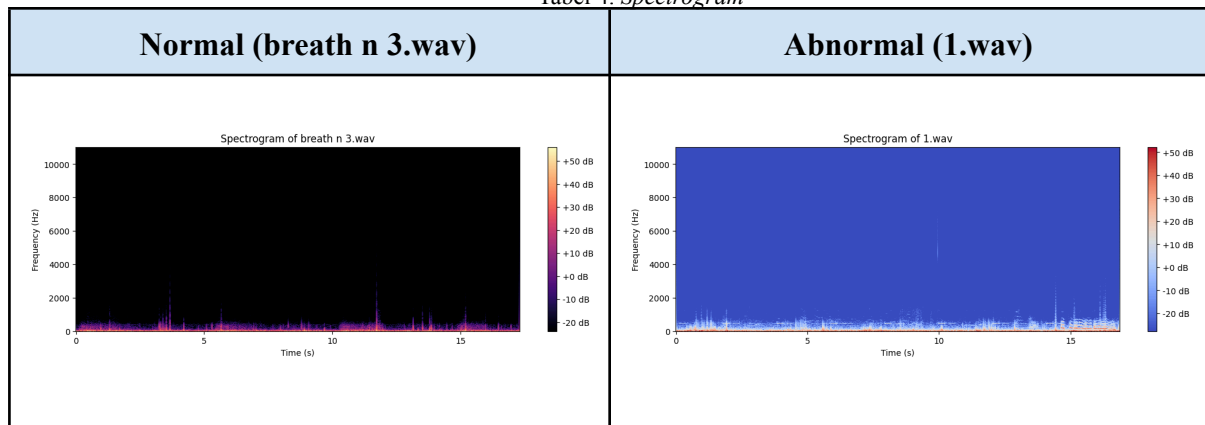
Tabel 3. *Spectrum*



*Spectrum* "normal" menunjukkan dominasi energi pada frekuensi rendah (0–1000 Hz) dengan penurunan signifikan di frekuensi tinggi, mencerminkan suara manusia yang lebih kaya di frekuensi rendah. Sebaliknya, *spectrum* "abnormal" menggambarkan sinyal dengan karakteristik frekuensi-waktu yang tidak teratur, dengan intensitas yang bervariasi pada berbagai frekuensi, mengindikasikan gangguan pada saluran pernapasan.

### 4.1.4 Spectrogram

Tabel 4. *Spectrogram*

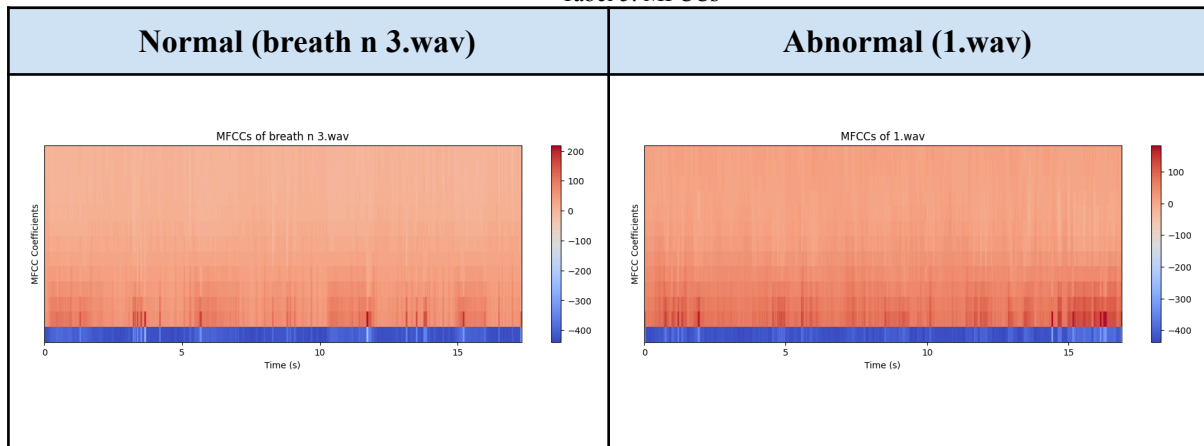


*Spectrogram* "normal" menunjukkan pola teratur dengan rentang frekuensi yang luas dan variasi intensitas yang sesuai dengan siklus pernapasan, mencerminkan kondisi pernapasan yang sehat. Sebaliknya, *spectrogram* "abnormal" menampilkan intensitas tinggi yang lebih dominan pada frekuensi rendah (di bawah 2000 Hz), sering terkait dengan gangguan seperti "*wheezing*", dengan variasi intensitas yang mengindikasikan gangguan pernapasan.



### 4.1.5 MFCC

Tabel 5. MFCCs



MFCC "normal" menunjukkan pola suara pernapasan dengan amplitudo yang lebih stabil, dengan warna hangat (merah, oranye) untuk amplitudo tinggi dan warna dingin (biru) untuk amplitudo rendah, mencerminkan pernapasan yang teratur. MFCC "abnormal" menunjukkan variasi yang lebih signifikan dalam warna, dengan perubahan pola amplitudo yang mengindikasikan gangguan pada pernapasan, seperti suara lebih keras (merah) atau lebih lembut (biru).

## 4.2 Modelling

### 4.2.1 Random Forest

Tabel 6. Hasil Evaluasi *Random Forest*

Feature Extraction	Accuracy	Precision	Recall	F1-Score
MFCC	0.89	0.40	0.50	0.44
Spectral Roll-Off	0.84	0.20	0.25	0.22
Zero-Crossing Rate	0.84	0.20	0.25	0.22
MFCC + Spectral Roll-Off	0.93	1	0.25	0.40
MFCC + Zero-Crossing Rate	0.93	1	0.25	0.40
Spectral Roll-Off + Zero-Crossing Rate	0.93	1	0.25	0.40
MFCC + Spectral Roll-Off + Zero-Crossing Rate	0.93	1	0.25	0.40

Pada tabel di atas, terlihat bagaimana model *Random Forest* bekerja dalam melakukan klasifikasi. Seluruh *feature extraction* pada model ini berhasil melakukan prediksi pada kelas negatif. Model dengan *feature extraction* MFCC + *Spectral Roll-Off* + *Zero-Crossing Rate* memiliki nilai metrik paling tinggi dengan akurasi sebesar 0.93, *precision* sebesar 1, *recall* sebesar 0.25, dan *F1-Score* sebesar 0.40.

#### 4.2.2 Light Gradient Boost Machine (LGBM)

Tabel 7. Hasil Evaluasi *Light Gradient Boost Machine*

Feature Extraction	Accuracy	Precision	Recall	F1-Score
MFCC	0.89	0.40	0.50	0.44
Spectral Roll-Off	0.69	0.14	0.50	0.22
Zero-Crossing Rate	0.80	0.22	0.50	0.31
MFCC + Spectral Roll-Off	0.87	0.38	0.75	0.50
MFCC + Zero-Crossing Rate	0.87	0.38	0.75	0.50
Spectral Roll-Off + Zero-Crossing Rate	0.80	0.22	0.50	0.31
MFCC + Spectral Roll-Off + Zero-Crossing Rate	0.91	0.50	0.75	0.60

Pada tabel di atas, terlihat bagaimana model *Light Gradient Boost Machine* (LGBM) bekerja dalam melakukan klasifikasi. Seluruh *feature extraction* pada model ini berhasil melakukan prediksi pada kelas negatif. Model dengan *feature extraction* MFCC + *Spectral Roll-Off* + *Zero-Crossing Rate* memiliki nilai metrik paling tinggi dengan akurasi sebesar 0.91, *precision* sebesar 0.50, *recall* sebesar 0.75, dan *F1-Score* sebesar 0.60.

#### 4.2.3 Support Vector Machine (SVM)

Tabel 8. Hasil Evaluasi *Support Vector Machine*

Feature Extraction	Accuracy	Precision	Recall	F1-Score
MFCC	0.42	0.13	1	0.24
Spectral Roll-Off	0.89	0.33	0.25	0.29
Zero-Crossing Rate	0.64	0.17	0.75	0.27
MFCC + Spectral Roll-Off	0.89	0.33	0.25	0.49
MFCC + Zero-Crossing Rate	0.42	0.13	1	0.24
Spectral Roll-Off + Zero-Crossing Rate	0.87	0.25	0.25	0.25
MFCC + Spectral Roll-Off + Zero-Crossing Rate	0.89	0.33	0.25	0.29

Pada tabel di atas, terlihat bagaimana model *Support Vector Machine* (SVM) bekerja dalam melakukan klasifikasi. Seluruh *feature extraction* pada model ini berhasil melakukan prediksi pada kelas negatif. Model dengan *feature extraction* *Spectral Roll-Off* memiliki nilai metrik paling tinggi dengan akurasi sebesar 0.89, *precision* sebesar 0.33, *recall* sebesar 0.25, dan *F1-Score* sebesar 0.29.

#### 4.2.4 Voting Classifier

Tabel 9. Hasil Evaluasi *Voting Classifier*

Feature Extraction	Accuracy	Precision	Recall	F1-Score
MFCC	0.89	0.33	0.25	0.29
Spectral Roll-Off	0.91	0.50	0.25	0.33
Zero-Crossing Rate	0.91	0	0	0
MFCC + Spectral Roll-Off	0.91	0.50	0.25	0.33
MFCC + Zero-Crossing Rate	0.89	0.33	0.25	0.29
Spectral Roll-Off + Zero-Crossing Rate	0.93	1	0.25	0.40
MFCC + Spectral Roll-Off + Zero-Crossing Rate	0.91	0.50	0.25	0.33

Pada tabel di atas, terlihat bagaimana model *Voting Classifier* bekerja dalam melakukan klasifikasi. Seluruh *feature extraction* pada model ini berhasil melakukan prediksi pada kelas negatif. Model dengan *feature extraction* *Spectral Roll-Off* +

*Zero-Crossing Rate* memiliki nilai metrik paling tinggi dengan akurasi sebesar 0.93, *precision* sebesar 1, *recall* sebesar 0.25, dan *F1-Score* sebesar 0.40.

#### 4.2.5 Pemilihan Model Terbaik

Tabel 10. Pemilihan Model Terbaik

Model	Feature Extraction	Accuracy	Precision	Recall	F1-Score
Random Forest	MFCC + Spectral Roll-Off + Zero-Crossing Rate	0.93	1	0.25	0.40
Light Gradient Boost Machine	MFCC + Spectral Roll-Off + Zero-Crossing Rate	0.91	0.50	0.75	0.60
Support Vector Machine	Spectral Roll-Off	0.89	0.33	0.25	0.29
Voting Classifier	Spectral Roll-Off + Zero-Crossing Rate	0.93	1	0.25	0.40

Setelah melakukan pemodelan dengan berbagai algoritma, dapat ditentukan model terbaik berdasarkan metrik-metrik yang digunakan untuk evaluasi performa. Tabel di atas menunjukkan bagaimana model *Random Forest* bekerja dalam melakukan klasifikasi. Seluruh *feature extraction* pada model ini berhasil memprediksi kelas negatif dengan baik. Model dengan kombinasi *feature extraction* MFCC + *Spectral Roll-Off* + *Zero-Crossing Rate* memberikan nilai akurasi tertinggi, yaitu 0.93, dengan *precision* sebesar 1, *recall* sebesar 0.25, dan *F1-Score* sebesar 0.40.

#### 4.3 Sistem Deteksi Penyakit Paru-Paru

Pada tabel di bawah ini, dapat dilihat hasil prediksi dari sistem deteksi menggunakan model terbaik yaitu *Random Forest* dengan kombinasi *feature extraction* MFCC + *Spectral Roll-Off* + *Zero-Crossing Rate* tersebut:

Tabel 11. Hasil Prediksi *Random Forest* dengan fitur MFCC + Spectral Roll-Off + Zero-Crossing Rate

Kelas	Prediksi Benar	Prediksi Salah	Jumlah File
Abnormal	197	0	197
Normal	22	3	25

Dari tabel di atas, dapat dilihat bahwa sistem deteksi berhasil memprediksi dengan benar 197 dari 197 file yang berlabel abnormal, tanpa ada prediksi salah pada kelas ini. Namun, pada kelas normal, terdapat 3 prediksi salah dari 25 file, yang menghasilkan *recall* rendah pada kelas tersebut. Meskipun demikian, hasil ini menunjukkan bahwa model lebih unggul dalam memprediksi kelas abnormal dibandingkan dengan kelas normal. Hal ini menunjukkan bahwa sistem ini memiliki kecenderungan untuk lebih berhati-hati dalam memprediksi normal, yang perlu diperhatikan jika deteksi kondisi normal juga sangat penting.

## BAB V

### KESIMPULAN DAN SARAN

#### 5.1 Kesimpulan

Berdasarkan analisis yang telah kami lakukan, diperoleh kesimpulan sebagai berikut:

- 1) Dalam tahap eksplorasi data (EDA), telah ditemukan bahwa dataset ini menunjukkan ketidakseimbangan yang signifikan antara kelas abnormal dan normal, dengan mayoritas data termasuk dalam kategori abnormal. Analisis gelombang suara (*waveform*), spectrum, spectrogram, dan MFCC menunjukkan perbedaan yang jelas antara kedua kelas tersebut, di mana kelas abnormal menunjukkan pola yang tidak teratur dan sering kali memiliki gangguan pada frekuensi rendah yang mengindikasikan potensi masalah pada saluran pernapasan. Hal ini memperkuat pentingnya menggunakan teknik ekstraksi fitur yang tepat untuk membedakan karakteristik antara kelas normal dan abnormal.
- 2) Pada tahap pemodelan, *Random Forest* dengan kombinasi fitur MFCC + *Spectral Roll-Off* + *Zero-Crossing Rate* menunjukkan hasil yang sangat baik dalam memprediksi kelas abnormal, dengan akurasi tertinggi (0.93) dan *precision* sebesar 1. Namun, *recall* untuk kelas normal tetap rendah, yang menunjukkan bahwa model lebih berhati-hati dalam memprediksi kelas normal. Model ini berhasil memprediksi kelas abnormal dengan sangat baik, tetapi perlu perbaikan untuk mengidentifikasi lebih banyak kasus normal dengan mengurangi prediksi salah pada kelas ini.

#### 5.2 Saran

Saran yang dapat kami berikan melalui penelitian ini adalah sebagai berikut:

- 1) Peningkatan model untuk kelas normal, karena *recall* yang rendah pada kelas normal menunjukkan perlunya perbaikan agar model lebih sensitif dalam mendeteksi kelas ini. Pendekatan seperti *balancing* data dengan *oversampling* atau *undersampling* dapat dipertimbangkan.
- 2) Dapat dilakukan percobaan menggunakan model-model lain, seperti *XGBoost* atau *Neural Networks*, yang dapat lebih sensitif terhadap data minoritas dan memberikan performa yang lebih baik dalam klasifikasi.
- 3) Peningkatan Fitur dan *Hyperparameter Tuning* dengan menambahkan fitur tambahan seperti *Chroma Feature* atau *Mel Spectrogram*, serta melakukan *fine-tuning hyperparameter* menggunakan teknik seperti *Grid Search* atau *Random Search* untuk meningkatkan akurasi model.

## DAFTAR PUSTAKA

- Baghel, N., Nangia, V. & Dutta, M.K. ALSD-Net: Automatic lung sounds diagnosis network from pulmonary signals. *Neural Comput & Applic* 33, 17103–17118 (2021). <https://doi.org/10.1007/s00521-021-06302-1>
- World Health Organization. (2024, November 6). Chronic obstructive pulmonary disease (COPD). Retrieved November 20, 2024, from [https://www.who.int/news-room/fact-sheets/detail/chronic-obstructive-pulmonary-disease-\(copd\)](https://www.who.int/news-room/fact-sheets/detail/chronic-obstructive-pulmonary-disease-(copd))
- [1] Septianto, Ryan Hendy. 2015. Diagnosa Penyakit Tanaman Kopi Arabika dengan Metode Modified K-Nearest Neighbor (MK-NN). Skripsi. Universitas Brawijaya, Malang.
- [2] Kumalasari, Noviana Ayu. 2014. Implementasi Algoritma Modified K-Nearest Neighbor (MKNN) untuk Menentukan Tingkat Resiko Penyakit Lemak Darah (Profil Lipid). Skripsi. Universitas Brawijaya, Malang.
- [3] S. Mujilawati, "Pre-Processing Text Mining Pada Data Twitter," Semin. Nas. Teknol. Inf. dan Komun., vol. 2016, no. Sentika, pp. 2089–9815, 2016.
- [4] V. W. Rumopa and J. Luther Mappadang, "Kontrol Penerangan Ruangan Menggunakan Sensor Suara ( Speech Recognition ) Berbasis Android," Tugas Akhir, 2015.
- [5] Fonseca, Eduardo, et al. "Freesound datasets: a platform for the creation of open audio datasets." *Hu X, Cunningham SJ, Turnbull D, Duan Z, editors. Proceedings of the 18th ISMIR Conference; 2017 oct 23-27; Suzhou, China.[Canada]: International Society for Music Information Retrieval; 2017. p. 486-93.. International Society for Music Information Retrieval (ISMIR), 2017.*
- [6] Swab, M. (2016). Mendeley data. *Journal of the Canadian Health Libraries Association/Journal de l'Association des bibliothèques de la santé du Canada*, 37(3).
- [7] Maulana, M. Aziz, Budi Darma Setiawan, and Rizal Setya Perdana. "Klasifikasi Kerusakan Permukaan Jalan Menggunakan Model MobileNetV3-Small." *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer* 8.8 (2024).
- [8] Antoniadis, A., Lambert-Lacroix, S., & Poggi, J. M. (2021). Random forests for global sensitivity analysis: A selective review. *Reliability Engineering & System Safety*, 206, 107312.
- [9] DPW Ellis, X Zeng, and McDermott JH, "Classifying Soundtracks With Audio Texture Features," ICASSP, 2011.
- [10] T Ishioka, "An Expansions of x-Means For Automatically Determining The Optimal Number of Clusters," in International Conference Computational Intelligence, pp. 91-96.
- [11] Zulfikar, Galih Shiddiq. *Klasifikasi kepribadian berdasarkan fitur audio spotify dan big five personality traits pada mahasiswa aktif prodi matematika menggunakan random forest*. BS thesis. Fakultas Sains dan Teknologi UIN Syarif Hidayatullah Jakarta.

- [12] Maulana, M. Aziz, Budi Darma Setiawan, and Rizal Setya Perdana. "Klasifikasi Kerusakan Permukaan Jalan Menggunakan Model MobileNetV3-Small." *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer* 8.8 (2024).
- [13] Antoniadis, A., Lambert-Lacroix, S., & Poggi, J. M. (2021). Random forests for global sensitivity analysis: A selective review. *Reliability Engineering & System Safety*, 206, 107312.
- [14] Aziz, R. M., Baluch, M. F., Patel, S., & Ganie, A. H. (2022). LGBM: a machine learning approach for Ethereum fraud detection. *International Journal of Information Technology*, 14(7), 3321-3331.
- [15] Fitriah Nur, Warsito Budi & D. A. I. Maruddani, "Analisis Sentimen GOJEK pada Media Sosial Twitter dengan Klasifikasi Support Vector Machine (SVM)", *Jurnal Gaussian*, vol. 9, no. 3, pp. 376 –390, 2020
- [16] O. H. Rahman, Abdillah Gunawan & Komarudin Agus, "Klasifikasi Ujaran Kebencian pada Media Sosial Twitter Menggunakan Support Vector Machine", *Jurnal Rekayasa Sistem dan Teknologi Informasi (RESTI)*, vol. 5, no. 1, pp. 17 –23, 2021
- [17] N. Agustina and C. N. Ihsan, "Pendekatan Ensemble untuk Analisis Sentimen Covid19 Menggunakan Pengklasifikasi Soft Voting," *Jurnal Teknologi Informasi dan Ilmu Komputer*, vol. 10, no. 2, p. 263, Apr. 2023, doi: 10.25126/jtiik.20231026215.