



# Evaluasi Kinerja Algoritma Klasifikasi dalam Memprediksi Gangguan Paru pada Dataset PEEP *Expiratory Occlusion*

## *Evaluation of Classification Algorithm Performance in Predicting Pulmonary Disorders on the PEEP Expiratory Occlusion Dataset*

Joelianto Darmawan<sup>1</sup>, Nurfira Ramadan<sup>2</sup>, Naura Mufliha Marhabani<sup>3</sup>, Siti Agrippina Alodia Yusuf<sup>4</sup>, Muhammad Imam Dinata<sup>5</sup>

<sup>1,2,3,4,5</sup> Sistem dan Teknologi Informasi, Universitas Muhammadiyah Mataram

darmawanjoe313@gmail.com, nurfiraramadannfrmdn@gmail.com, nauramufliha@gmail.com, siti.agrippina@ummat.ac.id, Imam.dinata@ummat.ac.id

**Abstract----** *Asthma is a chronic inflammatory disease of the airways that poses a significant global health concern, with a notable prevalence in Indonesia, affecting approximately 2.6% of the total population. The high risk and the often subjective nature of conventional diagnosis necessitate the development of objective and accurate early detection systems. This research focuses on analyzing and comparing the performance of three prominent Machine Learning algorithms Random Forest (RF), Support Vector Machine (SVM), and K-Nearest Neighbor (KNN) in classifying lung conditions, specifically Asthma versus Non-Asthma, utilizing features extracted from patient respiratory signals. The dataset employed is the Respiratory Dataset from PEEP Study with Expiratory Occlusion, comprising 78 subjects equally divided into 39 asthma patients and 39 normal subjects. The data underwent feature extraction, incorporating crucial respiratory parameters such as pressure, flow, and tidal volume, before being split into training (80%) and testing (20%) sets. Model performance was comprehensively evaluated using Accuracy, Precision, Recall (Sensitivity), and F1-Score metrics to ensure diagnostic reliability. Testing results demonstrate that the Random Forest algorithm delivered the most outstanding and stable performance across all metrics, achieving an Accuracy of 99,39%, F1-Score of 98,97%, Recall of 98,82%, and Precision of 99,13%. This was followed by KNN with an Accuracy of 90,97%. In contrast, SVM with the default RBF kernel showed the lowest performance, yielding an Accuracy of only 70,82% and a very low F1-Score of 2,56%, indicating a significant failure in identifying most positive asthma cases (low Recall). In conclusion, the Random Forest ensemble learning approach proves to be the most effective and reliable method for handling the complex and variable characteristics of respiratory physiological data, making it the recommended model for developing a robust early diagnostic system for asthma.*

**Keywords:** Asthma, Machine Learning, Random Forest (RF), Support Vector Machine (SVM), K-Nearest Neighbor (KNN)

**Abstrak----** *Asma merupakan penyakit inflamasi kronis saluran napas yang menjadi masalah kesehatan global, dengan prevalensi signifikan di Indonesia, mencapai sekitar 2,6% dari total populasi. Tingginya risiko dan sifat diagnosis konvensional yang seringkali subjektif mendorong kebutuhan mendesak akan sistem deteksi dini yang objektif dan akurat. Penelitian ini berfokus pada analisis dan perbandingan kinerja tiga algoritma Machine Learning Random Forest (RF), Support Vector Machine (SVM), dan K-Nearest Neighbor (KNN) dalam mengklasifikasikan kondisi paru, yaitu Asma versus Tidak Asma, berdasarkan fitur-fitur yang diekstraksi dari sinyal pernapasan pasien. Dataset yang digunakan berasal dari Respiratory Dataset from PEEP Study with Expiratory Occlusion, melibatkan 78 subjek yang*

\*Joelianto Darmawan

Email: Darmawanjoe313@gmail.com

terbagi rata menjadi 39 pasien asma dan 39 subjek normal. Data diproses melalui ekstraksi fitur yang mencakup parameter respirasi penting seperti pressure, flow, dan tidal volume, kemudian dibagi menjadi data latih (80%) dan data uji (20%). Kinerja model dievaluasi secara komprehensif menggunakan metrik Accuracy, Precision, Recall (Sensitivitas), dan F1-Score untuk memastikan keandalan prediksi. Hasil pengujian menunjukkan bahwa algoritma Random Forest mencatatkan performa paling unggul dan stabil di seluruh metrik, mencapai Akurasi 99,39%, F1-Score 98,97%, Recall 98,82%, dan Presisi 99,13%. Disusul oleh KNN dengan Akurasi 90,97%, sementara SVM dengan kernel RBF menunjukkan kinerja terendah, hanya mencapai Akurasi 70,82% dan F1-Score yang sangat rendah, yaitu 2,56%, mengindikasikan kegagalan dalam mengidentifikasi sebagian besar kasus asma positif. Kesimpulannya, pendekatan ensemble learning Random Forest terbukti paling efektif dalam menangani karakteristik kompleks dan bervariasi dari data fisiologis pernapasan, menjadikannya model yang direkomendasikan untuk pengembangan sistem diagnosis dini penyakit asma yang andal.

Kata Kunci: Asma, Machine Learning, Random Forest (RF), Support Vector Machine (SVM), K-Nearest Neighbor (KNN)

## I. PENDAHULUAN

Kesehatan pernapasan merupakan salah satu aspek penting dalam menjaga kualitas hidup manusia, di mana penyakit asma menjadi salah satu gangguan kronis yang memberikan dampak signifikan terhadap kesehatan masyarakat global. Asma merupakan penyakit inflamasi kronik pada saluran napas yang ditandai dengan gejala sesak napas, batuk, mengi, dan kekambuhan yang bersifat episodik. Menurut World Health Organization (WHO), jumlah penderita asma di seluruh dunia diperkirakan mencapai 250–262 juta orang, menjadikannya salah satu penyakit tidak menular yang paling umum di dunia [1]. Laporan *Global Burden of Asthma* menunjukkan bahwa prevalensi asma secara global bervariasi antar benua, dengan angka rata-rata sekitar 3,44% di Asia, 3,67% di Afrika, dan lebih tinggi di wilayah Amerika Utara serta Oseania [2]. Di Indonesia, asma masih menjadi salah satu penyakit kronik yang banyak dijumpai di masyarakat dan memberikan beban kesehatan yang cukup besar. Berdasarkan laporan VOI memperkirakan bahwa jumlah penderita asma di Indonesia mencapai 12 juta orang, atau sekitar 4,5% dari populasi nasional [3]. Beberapa faktor utama yang memicu kekambuhan asma di Indonesia antara lain paparan debu, asap rokok aktif maupun pasif, serta perubahan suhu atau udara dingin [4]. Kendala lain yang masih dihadapi adalah keterbatasan fasilitas diagnostik dan akses terhadap pengobatan di beberapa daerah, sehingga banyak kasus asma yang belum terdeteksi atau belum tertangani secara optimal. Oleh karena itu, upaya pencegahan, deteksi dini, serta peningkatan pengetahuan masyarakat mengenai pengelolaan asma menjadi hal yang sangat penting dalam menekan angka kejadian penyakit ini di Indonesia.

Pemanfaatan *machine learning* untuk prediksi kesehatan telah banyak diteliti, terutama pada klasifikasi berbagai penyakit. Pada penelitian yang dilakukan oleh [5], algoritma Random Forest digunakan untuk membantu mengambil keputusan pada sistem prediksi kanker. Penelitian lain yang dilakukan oleh [6] memprediksi dan mengklasifikasikan penyakit kanker menghasilkan akurasi sebesar 92,46% dengan menerapkan random forest dan metode optimasi fitur forward selection.

Pada penelitian ini, analisis kinerja dari metode klasifikasi pada klasifikasi kondisi paru normal dan tidak normal akan dilakukan. Tiga algoritma klasifikasi akan dibandingkan, yaitu algoritma Random Forest, Support Vector Machine (SVM) dan K-Nearest Neighbors (KNN). Dataset yang digunakan berasal dari Respiratory dataset from PEEP study with expiratory occlusion yang terdiri dari 78 subjek, dimana 39 subjek menderita asma dan 39 lainnya tidak menderita asma.

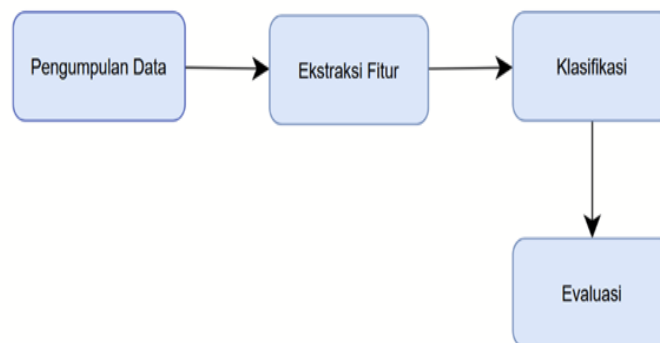
## II. LANDASAN TEORI

Perkembangan teknologi *Machine Learning* (ML) menawarkan solusi objektif dan efisien untuk mengatasi keterbatasan diagnosis konvensional. Pendekatan ML semakin banyak diimplementasikan di dunia medis karena kemampuannya dalam memproses dan mengklasifikasikan data secara cepat, menjadikannya alat bantu yang efektif dalam proses deteksi dini [9]. Klasifikasi adalah salah satu teknik data mining yang penting, bertujuan untuk memprediksi kategori atau kelas dari suatu objek, seperti mengidentifikasi apakah seorang pasien menderita penyakit (Asma) atau tidak [10]. Dengan membandingkan kinerja berbagai algoritma klasifikasi, penelitian dapat mengoptimalkan sensitivitas dan akurasi, yang sangat penting untuk meminimalkan false negative dalam skrining penyakit [9].

Penelitian mengenai penerapan algoritma *machine learning* dalam diagnosis penyakit paru telah berkembang pesat dalam beberapa tahun terakhir. Tujuan utamanya adalah membantu tenaga medis dalam mendeteksi penyakit paru seperti asma, pneumonia, atau kanker paru secara otomatis melalui analisis sinyal pernapasan, citra, maupun data klinis pasien. Pada penelitian sebelumnya yang dilakukan oleh [7], mendapat hasil bahwa algoritma Random Forest memiliki akurasi tertinggi dan stabilitas prediksi yang baik dibandingkan algoritma lainnya, menjadikannya model potensial untuk sistem pendukung keputusan medis dengan membandingkan beberapa algoritma seperti Random Forest, Support Vector Machine (SVM), dan Decision Tree dalam klasifikasi kondisi asma berdasarkan data klinis dan suara pernapasan. Kemudian penelitian yang dilakukan oleh [8], algoritma Random Forest mendapatkan nilai akurasi tertinggi, yaitu 92% dalam klasifikasi suara paru normal dan abnormal dengan mengimplementasikan model hibrid yang menggabungkan Support Vector Machine(SVM), Random Forest, dan Deep Neural Network. Penelitian ini menegaskan pentingnya kombinasi fitur audio dan model ensemble dalam meningkatkan akurasi diagnosis penyakit paru.

## III. METODE PENELITIAN

Pada penelitian ini, terdapat empat langkah yang digunakan dalam klasifikasi penyakit paru, yaitu pengumpulan data, ekstraksi fitur, klasifikasi, dan evaluasi, seperti yang ditunjukkan pada Gambar 1. Masing-masing tahapan akan dijelaskan pada sub bab berikut.



Gambar 1. Alur Penelitian

### A. Pengumpulan Data

Data yang digunakan dalam penelitian ini merupakan dataset yang berasal dari Physionet [11], yang terdiri dari 80 subjek laki-laki dan perempuan. Berkas disimpan berdasarkan nomor subjek ('01' hingga '80'). Berkas berisi data berikut yang diambil sampelnya pada 100Hz: tekanan gauge(gauge pressure); tekanan diferensial inspirasi(inspiratory differential pressure); tekanan diferensial ekspirasi(expiratory differential pressure); lingkaran dada(chest circumference); lingkaran

perut; kedalaman awal penampang dada; dan lebar awal penampang dada. Dataset berupa data *time series* dengan jumlah data untuk setiap subjek 111.000 hingga 119.000 dengan label 0 untuk kondisi paru normal, dan 1 untuk kondisi paru asma. Pada penelitian ini, hanya 78 subjek yang akan digunakan, dan 100.000 data untuk masing-masing subjek, sehingga total dataset yang digunakan sejumlah 7.800.000.

### B. Ekstraksi fitur

Pada tahap ini, rata-rata (*mean*) fitur dari masing-masing dataset, yaitu pressure [*cmH<sub>2</sub>O*], flow [*L/s*], tidal volume [*L*], chest circumference [*mm*], abdominal circumference [*mm*], indeks inspirasi, aerasi global, dan sifat indeks inspirasi data aerasi dihitung dengan pergeseran window sebesar 200, sehingga jumlah dataset yang diperoleh per subjek sebesar 500. Total dataset dari keseluruhan subjek berjumlah 39.000. Fitur-fitur yang digunakan pada penelitian ini dipilih berdasarkan studi sebelumnya [12], dimana fitur-fitur tersebut mengindikasikan adanya asma atau tidak.

### C. Klasifikasi

Proses klasifikasi adalah proses dalam *machine learning* yang bertujuan untuk mengelompokkan data ke dalam kategori yang telah ditentukan berdasarkan pola fitur yang telah dipelajari dari data pelatihan [13]. Pada tahap ini, terdapat tiga algoritma klasifikasi yang akan dibandingkan, yaitu RF, SVM, dan KNN. Masing-masing algoritma tersebut dijelaskan pada sub bab berikut.

#### 1. Random Forest (RF)

Arsitektur Random Forest dibangun atas dasar konsep *ensemble learning*, di mana sejumlah *decision tree* digabungkan untuk menghasilkan prediksi yang lebih akurat dan stabil [14]. Proses pelatihannya diawali dengan bootstrap sampling, yaitu pengambilan subset data pelatihan secara acak dengan pengembalian. Setiap subset digunakan untuk melatih satu *decision tree* yang berbeda, dan di setiap node pohon, hanya sebagian fitur yang dipilih secara acak (*feature bagging*) untuk menciptakan keragaman antar pohon. Setelah semua pohon selesai dibangun, prediksi dilakukan dengan menggabungkan hasil dari seluruh pohon menggunakan mekanisme voting mayoritas untuk klasifikasi atau averaging untuk regresi, dengan rumus umum:

$$\hat{y} = \text{mode}\{h_1(x), h_2(x), \dots, h_n(x)\} \quad (1)$$

Dimana  $h_i(x)$  adalah hasil prediksi dari pohon ke- $i$  [14].

Pada penelitian ini, algoritma Random Forest menggunakan jumlah pohon keputusan sebanyak 100 trees yang merupakan nilai default dari parameter `n_estimators` pada pustaka *Scikit-learn*. Jumlah tersebut dipilih karena telah terbukti memberikan keseimbangan antara akurasi dan efisiensi komputasi pada berbagai studi klasifikasi medis [15]. Parameter `random_state=42` digunakan untuk menjaga konsistensi hasil pelatihan dan memastikan reproduktibilitas model.

#### 2. Support Vector Machine (SVM)

Arsitektur *Support Vector Machine* (SVM) berfokus pada pencarian *hyperplane* optimal yang memisahkan data dari dua kelas atau lebih dengan margin maksimum [16]. SVM bekerja dengan prinsip mencari batas keputusan (*decision boundary*) yang mampu mengoptimalkan jarak antara titik data dari kelas yang berbeda. Apabila data tidak dapat dipisahkan secara linear, SVM menggunakan fungsi *kernel*, seperti *Radial Basis Function* (RBF), untuk memetakan data ke ruang fitur berdimensi lebih tinggi sehingga dapat dipisahkan secara linear. Pemilihan kernel dan pengaturan parameter, khususnya konstanta regulasi (C) dan parameter  $\gamma$  (*gamma*), berperan penting dalam menentukan kinerja model. Rumus umum SVM dapat ditulis sebagai:

$$f(x) = w^T x + b \quad (2)$$

dengan  $w$  adalah vektor bobot,  $T$  adalah transpose dari  $w$  sehingga menjadi  $1 \times n$ ,  $x$  adalah vektor fitur, dan  $b$  adalah bias.

Pada penelitian ini, algoritma Support Vector Machine (SVM) menggunakan fungsi kernel Radial Basis Function (RBF) untuk memetakan data ke ruang fitur berdimensi lebih tinggi agar pemisahan antar kelas dapat dilakukan secara linear. Parameter yang digunakan pada model adalah  $C = 1.0$  dan  $\gamma = \text{'scale'}$ , yang merupakan nilai default pada pustaka *Scikit-learn*. Parameter  $C$  berfungsi sebagai faktor regularisasi yang mengontrol keseimbangan antara margin pemisahan yang lebar dan kesalahan klasifikasi, sedangkan  $\gamma$  menentukan tingkat pengaruh suatu titik data terhadap pembuatan batas keputusan. Pengaturan nilai ini memungkinkan model untuk menangkap pola non-linear dalam data respirasi tanpa risiko overfitting yang berlebihan [15].

### 3. K-Nearest Neighbors (K-NN)

Arsitektur K-Nearest Neighbor (KNN) termasuk kategori *lazy learning*, karena tidak memerlukan proses pelatihan model secara eksplisit. KNN bekerja dengan menyimpan seluruh data pelatihan dan menentukan kelas data baru berdasarkan kedekatan jarak terhadap data pelatihan yang sudah diketahui kelasnya [17]. Algoritma ini sangat bergantung pada pemilihan metrik jarak dan parameter  $K$ , yaitu jumlah tetangga terdekat yang dipertimbangkan dalam proses klasifikasi.

Jarak antara dua titik data  $x_i$  dan  $x_j$  dapat dihitung menggunakan rumus Euclidean distance:

$$d(x_i, x_j) = \sqrt{\sum_{k=1}^n (x_{ik} - x_{jk})^2} \quad (3)$$

Pada penelitian ini, algoritma K-Nearest Neighbor (KNN) menggunakan jumlah tetangga terdekat ( $K = 3$ ) dalam proses klasifikasi. Nilai  $K$  tersebut dipilih untuk menjaga keseimbangan antara sensitivitas terhadap noise dan kemampuan generalisasi model. Dengan  $K$  yang relatif kecil, model mampu menangkap pola lokal antar data dengan baik tanpa kehilangan akurasi akibat pengaruh data yang terlalu jauh. Pengukuran jarak antar data dilakukan menggunakan metrik Euclidean Distance, yang merupakan metode umum untuk mengukur kemiripan antar titik dalam ruang fitur berdimensi kontinu. Nilai  $K = 3$  juga telah terbukti memberikan hasil yang stabil dalam berbagai studi klasifikasi penyakit berbasis data medis menggunakan KNN, seperti penelitian yang dilakukan oleh [15], bahwa pemilihan parameter  $K$  yang optimal mampu meningkatkan akurasi model secara signifikan dalam prediksi penyakit menggunakan data fisiologis.

### D. Evaluasi

Evaluasi model merupakan tahapan untuk menganalisis dan mengukur seberapa efektif dan andal model yang telah dilatih dalam membuat prediksi yang tepat pada data uji yang belum pernah dilihat sebelumnya [18]. Pada penelitian ini, terdapat 4 pengukuran yang akan digunakan, yaitu Presisi, Recall, Akurasi, dan F1-Score.

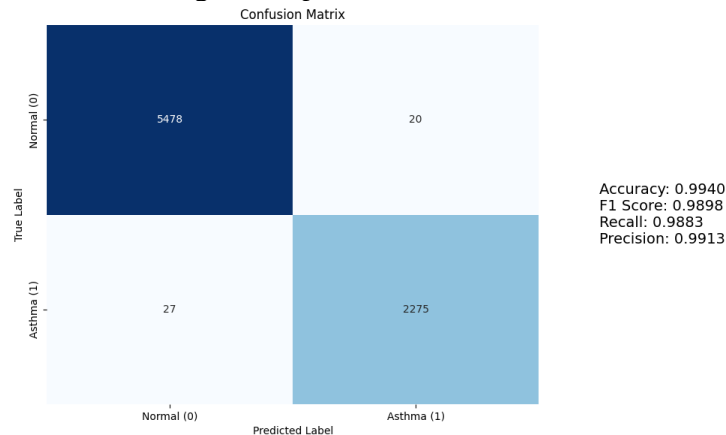
## IV. HASIL / IMPLEMENTASI MODEL DAN PEMBAHASAN

### A. Hasil Klasifikasi

#### 1. Random Forest

Model Random Forest menunjukkan performa terbaik dalam klasifikasi penyakit paru. Berdasarkan hasil pengujian, diperoleh nilai akurasi sebesar 99,39%, F1-score 98,97%, recall

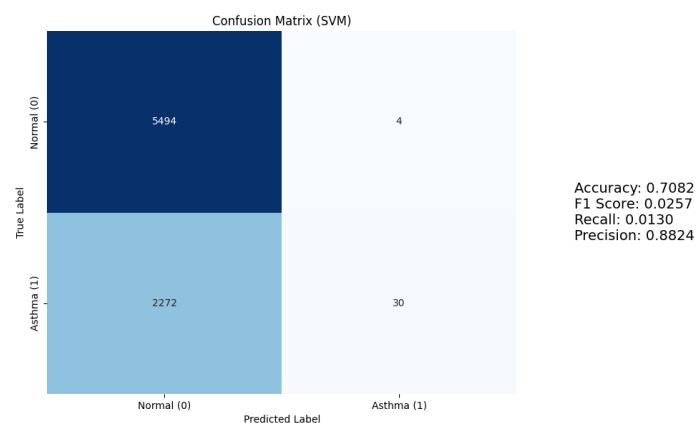
98,82% dan presisi 99,13%. Gambar 2 menunjukkan hasil pengenalan dari Random Forest, dapat dilihat bahwa hampir seluruh data uji berhasil diklasifikasikan dengan benar, dengan hanya satu atau dua kesalahan prediksi dari total sampel. Kinerja tinggi ini menunjukkan bahwa Random Forest mampu mempelajari pola kompleks antara fitur-fitur respirasi dan status kesehatan paru secara efektif. Penggunaan banyak pohon keputusan (*ensemble of trees*) membuat model lebih stabil dan tidak mudah overfitting terhadap data latih.



Gambar 2. Hasil klasifikasi Random Forest

### B. Support Vector Machine (SVM)

Model Support Vector Machine (SVM) dengan kernel Radial Basis Function (RBF) menghasilkan akurasi sebesar 70,82%, F1-score 2,56%, recall 1,30% dan presisi 88,24%. Hasil ini menunjukkan bahwa model SVM kurang optimal dalam mengklasifikasikan data respirasi menjadi dua kelas yang berbeda. Rendahnya nilai F1-score dan recall menandakan bahwa model gagal mengenali sebagian besar kasus asma (kelas positif), meskipun akurasi keseluruhan masih di atas 70%. Hal ini disebabkan oleh distribusi data yang tidak sepenuhnya linear serta kurang optimalnya parameter kernel (C dan gamma), sebagaimana dibahas secara mendalam dalam literatur terkini bahwa parameter yang tidak tepat dapat menurunkan sensitivitas dan f1-score model [20]. Dalam konteks medis, performa seperti ini menunjukkan bahwa SVM belum cukup andal untuk diagnosis penyakit paru tanpa optimasi parameter lebih lanjut.

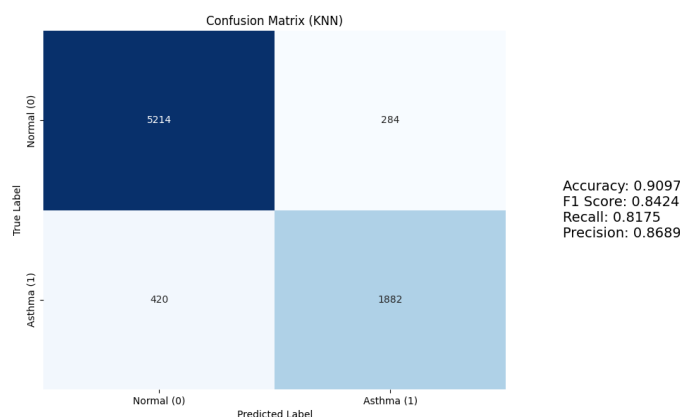


Gambar 3. Hasil klasifikasi Support Vector Machine (SVM)

### C. Hasil K-Nearest Neighbor (KNN)

Pada algoritma KNN dengan parameter  $K = 3$ , diperoleh nilai akurasi sebesar 90,97%, F1-score 84,24%, recall 81,75% dan presisi 86,89%. Confusion matrix menunjukkan bahwa sebagian

besar data berhasil diklasifikasikan dengan benar, meskipun terdapat beberapa kesalahan pada kelas asma. Kinerja KNN yang cukup tinggi ini menunjukkan bahwa pendekatan berbasis jarak masih efektif untuk mendeteksi pola kemiripan antar pasien berdasarkan fitur-fitur sinyal pernapasan. Namun demikian, karena KNN termasuk *lazy learner*, waktu komputasi meningkat seiring bertambahnya jumlah data dan model rentan terhadap *noise*.



Gambar 4. Hasil Klasifikasi K-Nearest Neighbor (KNN)

Hasil klasifikasi keempat algoritma ditampilkan pada tabel 1 berikut:

TABEL 1. PERBANDINGAN KINERJA ALGORITMA

| Nama Algoritma               | Akurasi (%) | F1-Score (%) | Recall (%) | Presisi (%) |
|------------------------------|-------------|--------------|------------|-------------|
| Random Forest                | 99,39       | 98,97        | 98,82      | 99,13       |
| K-Nearest Neighbor (KNN)     | 90,97       | 84,24        | 81,75      | 86,89       |
| Support Vector Machine (SVM) | 70,82       | 2,56         | 1,30       | 88,24       |

Dari hasil tersebut dapat disimpulkan bahwa Random Forest memberikan kinerja paling unggul dan konsisten, disusul oleh KNN yang cukup kompetitif, sementara SVM memiliki performa terendah dalam kasus ini. Perbedaan signifikan antara algoritma tersebut menunjukkan pentingnya pemilihan metode yang sesuai dengan karakteristik data medis, khususnya yang bersifat non-linear dan berukuran kecil.

#### D. Pembahasan

Hasil pengujian terhadap tiga algoritma klasifikasi yaitu Random Forest, K-Nearest Neighbor (KNN), dan Support Vector Machine (SVM) menunjukkan adanya perbedaan signifikan dalam tingkat akurasi, kemampuan generalisasi, serta keandalan model dalam mengenali pola kondisi paru-paru antara pasien normal dan penderita asma. Perbedaan ini erat kaitannya dengan karakteristik algoritma masing-masing, cara pengambilan keputusan, serta sensitivitas terhadap distribusi data dan fitur yang digunakan.

Algoritma Random Forest menghasilkan performa terbaik dengan akurasi sebesar 99,39%, F1-score sebesar 98,97%, recall sebesar 98,82% dan presisi sebesar 99,13%. Hasil ini membuktikan bahwa metode *ensemble learning* yang menggabungkan banyak pohon keputusan mampu meminimalkan kesalahan prediksi dan mengatasi masalah *overfitting* yang sering terjadi pada model tunggal. Setiap pohon dalam Random Forest dibangun dari subset data dan fitur yang berbeda, sehingga model mampu menangkap variasi pola pada data fitur suara paru secara lebih

menyeluruh. Kemampuan ini sangat penting mengingat data fisiologis seperti sinyal paru memiliki kompleksitas tinggi dan rentan terhadap *noise*. Penelitian yang dilakukan oleh [21] menunjukkan bahwa Random Forest unggul dibandingkan SVM dan regresi logistik dalam memprediksi risiko asma berdasarkan fitur klinis dan lingkungan. Selain itu, [22] juga membuktikan bahwa Random Forest efektif untuk data medis yang kompleks dan bising karena kemampuannya dalam melakukan *noise reduction* dan *feature sampling* secara acak, sehingga meningkatkan stabilitas model. Dengan demikian, algoritma Random Forest terbukti menjadi metode yang paling andal untuk klasifikasi penyakit paru yang melibatkan fitur suara kompleks dan beragam.

Sementara itu, algoritma K-Nearest Neighbor (KNN) memberikan hasil yang cukup kompetitif dengan akurasi sebesar 90,97%, F1-score 84,24%, recall 81,75%, dan presisi 86,89%. KNN bekerja berdasarkan prinsip kemiripan jarak antar data, sehingga performanya sangat bergantung pada pemilihan nilai  $k$  dan skala fitur yang digunakan. Dengan distribusi data yang seimbang, KNN mampu memberikan hasil klasifikasi yang konsisten karena setiap kelas memiliki representasi yang proporsional dalam proses penentuan tetangga terdekat. Akurasi yang sedikit lebih rendah dibandingkan Random Forest dapat disebabkan oleh sensitivitas KNN terhadap *noise* dan keberadaan data dengan jarak antar kelas yang relatif berdekatan. Namun, hasil yang diperoleh tetap menunjukkan bahwa KNN efektif dalam mengenali pola utama antara kelas normal dan asma. Penelitian oleh [23] membuktikan bahwa KNN memiliki performa tinggi dalam klasifikasi data medis ketika dilakukan normalisasi fitur yang baik. Dengan demikian, KNN dapat dianggap sebagai metode yang andal untuk diagnosis penyakit paru berbasis suara ketika data memiliki representasi yang proporsional antar kelas.

Adapun algoritma Support Vector Machine (SVM) menunjukkan performa paling rendah dengan tingkat akurasi sebesar 70,82%. Hasil ini mengindikasikan bahwa model SVM dengan kernel Radial Basis Function (RBF) belum mampu menangani kompleksitas data non-linear yang terdapat pada fitur suara paru secara efektif. Rendahnya performa SVM diduga disebabkan oleh pemilihan parameter  $C$  sebesar 1.0 dan  $\gamma$  dengan nilai default *scale*, yang belum optimal terhadap karakteristik data. Nilai  $C$  yang tidak tepat dapat menyebabkan model terlalu menyesuaikan diri terhadap data pelatihan (*overfitting*) atau justru terlalu sederhana (*underfitting*), sedangkan  $\gamma$  yang tidak sesuai dapat membuat batas keputusan (*decision boundary*) menjadi terlalu halus atau terlalu kompleks. Kondisi ini menunjukkan bahwa proses *tuning parameter* belum dilakukan secara optimal, sehingga kemampuan model dalam membedakan pola suara paru menjadi terbatas. Temuan ini sejalan dengan penelitian oleh [24] yang menyatakan bahwa kombinasi antara  $C$  dan  $\gamma$  sangat memengaruhi performa SVM. Hal serupa juga dikemukakan oleh [25] yang meneliti efektivitas *parameter tuning* pada basis data patologi suara, di mana penyesuaian nilai  $C$  dan  $\gamma$  secara signifikan meningkatkan akurasi model SVM pada data non-linear.

Berdasarkan hasil keseluruhan, dapat disimpulkan bahwa pendekatan ensemble learning seperti Random Forest lebih efektif dalam menangani data dengan karakteristik kompleks dan variabilitas tinggi seperti data fisiologis paru. Hasil evaluasi juga menunjukkan bahwa pemilihan algoritma dan parameter yang sesuai sangat mempengaruhi hasil klasifikasi. Dengan demikian, tahap pra-pemrosesan data, pemilihan fitur relevan, serta penalaan parameter (*hyperparameter tuning*) merupakan aspek penting untuk meningkatkan akurasi model secara keseluruhan.

## V. KESIMPULAN

Berdasarkan hasil penelitian terhadap tiga algoritma klasifikasi, yaitu Random Forest, K-Nearest Neighbor (KNN), dan Support Vector Machine (SVM) dalam mendeteksi kondisi paru-paru (asma dan tidak asma) berdasarkan fitur suara pernapasan, diperoleh bahwa Random Forest menunjukkan performa terbaik dengan akurasi sebesar 99,39%, F1-score 98,97%, recall 98,82%,

dan presisi 99,13%. Hasil tersebut menegaskan bahwa pendekatan *ensemble learning* dalam penggabungan beberapa pohon keputusan dapat meningkatkan stabilitas model dan mengurangi risiko overfitting. Keunggulan Random Forest juga menunjukkan efektivitasnya dalam menangani data fisiologis yang memiliki karakteristik kompleks dan variabilitas tinggi. Sementara itu, algoritma K-Nearest Neighbor (KNN) menghasilkan performa yang cukup kompetitif dengan akurasi 90,97%, F1-score 84,24%, recall 81,75%, dan presisi 86,89%. Hasil ini menunjukkan bahwa KNN mampu mengenali pola kemiripan antar data dengan baik, terutama ketika dataset terdistribusi secara seimbang dan telah melalui proses normalisasi yang tepat. Adapun Support Vector Machine (SVM) mencatatkan performa paling rendah dengan akurasi 70,82%, yang mengindikasikan bahwa model dengan kernel Radial Basis Function (RBF) belum sepenuhnya mampu merepresentasikan kompleksitas hubungan non-linear pada fitur suara paru. Rendahnya performa ini juga disebabkan oleh proses tuning parameter (C dan gamma) yang belum optimal terhadap karakteristik data yang digunakan.

Secara keseluruhan, penelitian ini membuktikan bahwa pemilihan algoritma dan konfigurasi parameter yang tepat berperan penting dalam meningkatkan akurasi sistem klasifikasi berbasis data suara paru. Berdasarkan hasil pengujian, Random Forest direkomendasikan sebagai model yang paling andal untuk sistem pendukung keputusan diagnosis asma, karena kemampuannya mengatasi noise, menangkap pola non-linear, dan memberikan hasil prediksi yang konsisten.

Untuk penelitian selanjutnya, disarankan dilakukan optimasi hyperparameter menggunakan pendekatan sistematis seperti Grid Search, Bayesian Optimization, atau Genetic Algorithm untuk memperoleh kinerja model yang lebih optimal. Selain itu, integrasi metode feature selection dan deep feature extraction menggunakan teknik seperti Principal Component Analysis (PCA), Autoencoder, atau Convolutional Neural Network (CNN) dapat meningkatkan kualitas representasi fitur dari sinyal suara paru.

## REFERENSI

- [1] WHO, "Asthma." Accessed: Oct. 15, 2025. [Online]. Available: <https://www.who.int/news-room/fact-sheets/detail/asthma>
- [2] A. P. J. Rabe, W. J. Loke, K. Gurjar, A. Brackley, and D. E. Lucero-Prisno, "Global Burden of Asthma, and Its Impact on Specific Subgroups: Nasal Polyps, Allergic Rhinitis, Severe Asthma, Eosinophilic Asthma," *J. Asthma Allergy*, vol. 16, no. October, pp. 1097–1113, 2023, doi: 10.2147/JAA.S418145.
- [3] VOI, "Diagnosis Devices Not Even, 12 Million Indonesians Reportedly Suffer From Asma." Accessed: Oct. 15, 2025. [Online]. Available: <https://voi.id/en/news/276778>
- [4] W. Agustina, "Dust, Cigarette Smoke, and Cold Air are the Main Risk Factors for Asthma Attacks," *J. Respirologi Indones.*, vol. 44, no. 4, 2024, doi: 10.36497/jri.v44i4.647.
- [5] Arifin Yusuf Permana, Hari Noer Fazri, M.Fakhrizal Nur Athoilah, Mohammad Robi, and Ricky Firmansyah, "Penerapan Data Mining Dalam Analisis Prediksi Kanker Paru Menggunakan Algoritma Random Forest," *J. Ilm. Tek. Inform. dan Komun.*, vol. 3, no. 2, pp. 27–41, 2023, doi: 10.55606/juitik.v3i2.472.
- [6] L. Rangga, A. Tarigan, T. Informatika, R. Forest, O. Fitur, and F. Selection, "Optimlasasi Fitur Dengan Forward Selection Pada Estimasi Tingkat Penyakit Paru-Paru Menggunakan Algoritma," vol. 8, no. 5, pp. 10341–10348, 2024.
- [7] K. C. H. Tsang, H. Pinnock, A. M. Wilson, and S. A. Shah, "Application of Machine Learning Algorithms for Asthma Management with mHealth: A Clinical Review," *J. Asthma Allergy*, vol. 15, pp. 855–873, 2022, doi: 10.2147/JAA.S285742.
- [8] K. V. Suma, D. Koppad, P. Kumar, N. A. Kantikar, and S. Ramesh, "Multi-task Learning for Lung Sound and Lung Disease Classification," *SN Comput. Sci.*, vol. 6, no. 1, pp. 1–22, 2025, doi: 10.1007/s42979-024-03506-9.
- [9] N. Nuraeni and P. Astuti, "Pendekatan Machine Learning untuk Deteksi Dini Kanker Paru-Paru: Mengoptimalkan Sensitivitas dan Akurasi," *J. Inform. Polinema*, vol. 11, no. 3, pp. 339–346, 2025, doi: 10.33795/jip.v11i3.7011.
- [10] S. Rahmaeda and R. Prathivi, "Komparasi Metode SVM dan Logistic Regression untuk Klasifikasi Hipotesa Penyakit Kanker Paru Paru Berdasarkan Gejala Awal," *Kesatria J. Penerapan Sist. Inf. (Komputer dan Manajemen)*, vol. 6, no. 1, pp. 210–218, 2025, doi: 10.30645/kesatria.v6i1.562.

- [11] J. G. Guy, E. F. S.; Clifton, J. A.; Caljé-van der Klei, T.; Chen, R.; Knopp, J.; Moeller, K.; Chase, "Respiratory dataset from PEEP study with expiratory occlusion (version 1.0.0)," 2023. doi: 10.13026/d767-e709.
- [12] G. N. Beken, S. Z. Aydin, and B. Erkmén, "Deep Long Short-Term Memory Networks for Asthma Disease Prediction from Respiratory Dataset Obtained from the PEEP Expiratory Occlusion," *2024 47th Int. Conf. Telecommun. Signal Process. TSP 2024*, pp. 239–242, 2024, doi: 10.1109/TSP63128.2024.10605923.
- [13] T. Xia, J. Han, and C. Mascolo, "Exploring machine learning for audio-based respiratory condition screening: A concise review of databases, methods, and open issues," *Exp. Biol. Med.*, vol. 247, no. 22, pp. 2053–2061, 2022, doi: 10.1177/15353702221115428.
- [14] Z. Jin, J. Shang, Q. Zhu, C. Ling, W. Xie, and B. Qiang, "RFRSF: Employee Turnover Prediction Based on Random Forests and Survival Analysis," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 12343 LNCS, pp. 503–515, 2020, doi: 10.1007/978-3-030-62008-0\_35.
- [15] D. K. Barupal and O. Fiehn, "Generating the blood exposome database using a comprehensive text mining and database fusion approach," *Environ. Health Perspect.*, vol. 127, no. 9, pp. 2825–2830, 2019, doi: 10.1289/EHP4713.
- [16] L. Mishra, S. Verma, and S. Varma, "Brain Tumor Classification Using Feature Extraction and Non-linear SVM Hybrid Model," *Commun. Comput. Inf. Sci.*, vol. 1894 CCIS, pp. 3–14, 2023, doi: 10.1007/978-3-031-43145-6\_1.
- [17] R. K. Halder, M. N. Uddin, M. A. Uddin, S. Aryal, and A. Khraisat, "Enhancing K-nearest neighbor algorithm: a comprehensive review and performance analysis of modifications," *J. Big Data*, vol. 11, no. 1, 2024, doi: 10.1186/s40537-024-00973-y.
- [18] I. M. Esa Darmayasa Adi Putra, I. Fauzi, K. S. Krishna Prasad, K. S. Krishna Prasad, I. M. Putra Arya Winata, and I. W. Widhiada, "Evaluasi Model Machine Learning Klasifikasi Gerak Tangan Untuk Sistem Kontrol Prototipe Prostesis Tangan," *Maj. Ilm. Teknol. Elektro*, vol. 22, no. 1, p. 141, 2023, doi: 10.24843/mite.2023.v22i01.p18.
- [19] A. Zheng, A. Casari, and M. Edwards, *Feature engineering for machine learning: Principles and techniques*, vol. 53, no. 4. 2021.
- [20] K. Thurnhofer-hemsi, E. López-rubio, and M. A. Molina-cabello, "R s v m," 2020.
- [21] B. Belieffain, F. Daeli, and U. P. Sanjaya, "Comparative Analysis of Logistic Regression , Random Forest , and SVM for Asthma Risk Prediction Using Demographic , Clinical , and Environmental Features," vol. 9, no. 5, pp. 2187–2194, 2025.
- [22] Y. Li, "Research on ECG Classification Algorithm Based on Wavelet Transform Noise Reduction and Random Forest," *Appl. Comput. Eng.*, vol. 146, no. 1, pp. 33–42, 2025, doi: 10.54254/2755-2721/2025.tj21924.
- [23] M. M. A. H. L. F. Lidimillah, "Implementation of the K-Nearest Neighbors Method for Determining Fetal Health Status," *J. Tek. Inform.*, vol. 5, no. 4, pp. 329–336, 2024, doi: 10.52436/1.jutif.2024.5.4.2173.
- [24] C.-J. L. Chih-Wei Hsu, Chih-Chung Chang, "A Practical Guide to Support Vector Classification," *BJU Int.*, vol. 101, no. 1, pp. 1396–1400, 2008, [Online]. Available: <http://www.csie.ntu.edu.tw/%7B~%7Dcjlin/papers/guide/guide.pdf>
- [25] Nurul Fariesya Suhaila Md Sazihan, N. M. Abdul Latiff, Nik Noordini Nik Abd Malik, Syamsiah Mashohor, and Fahad Taha Al-Dhief, "Evaluating the Effectiveness of Parameter Tuning for Support Vector Machine on Voice Pathology Database," *Elektr. J. Electr. Eng.*, vol. 24, no. 2, pp. 118–124, 2025, doi: 10.11113/elektrika.v24n2.628.