

DETEKSI MALWARE PADA FILE PDF DENGAN ALGORITMA SVM DAN RANDOM FOREST

Tugas Akhir

**Diajukan untuk Memenuhi Salah Satu Persyaratan Mencapai Derajat
Sarjana Komputer**



Dibuat Oleh:

Kendrick Chandra

200710554

**PROGRAM STUDI INFORMATIKA
FAKULTAS TEKNOLOGI INDUSTRI
UNIVERSITAS ATMA JAYA YOGYAKARTA
2025**

HALAMAN PENGESAHAN

Tugas Akhir Berjudul

DETEKSI MALWARE PADA FILE PDF DENGAN ALGORITMA SVM DAN RANDOM FOREST

yang disusun oleh

Kendrick Chandra

200710554

dinyatakan telah memenuhi syarat pada tanggal 24 Juni 2025

		Keterangan
Dosen Pembimbing 1	: Paulus Mudjihartono, S.T.,M.T., Ph. D	Telah Menyetujui
Dosen Pembimbing 2	: Vinindita Citrayasa, S.Pd., M.Hum	Telah Menyetujui
Tim Penguji		
Penguji 1	: Paulus Mudjihartono, S.T.,M.T., Ph. D	Telah Menyetujui
Penguji 2	: Wilfridus Bambang Triadi H, ST., M.Cs	Telah Menyetujui
Penguji 3	: Prof. Dr. Ir. Alb. Joko Santoso, M.T.	Telah Menyetujui

Yogyakarta, 24 Juni 2025

Universitas Atma Jaya Yogyakarta

Teknologi Industri

Dekan

ttd.

Dr. Ir. Parama Kartika Dewa SP., S.T., M.T.

Dokumen ini merupakan dokumen resmi UAJY yang tidak memerlukan tanda tangan karena dihasilkan secara elektronik oleh Sistem Bimbingan UAJY. UAJY bertanggung jawab penuh atas informasi yang tertera di dalam dokumen ini

PERNYATAAN ORISINALITAS & PUBLIKASI ILMIAH

Saya yang bertanda tangan di bawah ini:

Nama Lengkap : Kendrick Chandra
NPM : 200710554
Program Studi : Informatika
Fakultas : Teknologi Industri
Judul Penelitian : Deteksi Malware pada File PDF dengan Algoritma
SVM dan Random Forest

Menyatakan dengan ini:

1. Tugas Akhir ini adalah benar tidak merupakan salinan sebagian atau keseluruhan dari karya penelitian lain.
2. Memberikan kepada Universitas Atma Jaya Yogyakarta atas penelitian ini, berupa Hak untuk menyimpan, mengelola, mendistribusikan, dan menampilkan hasil penelitian selama tetap mencantumkan nama penulis.
3. Bersedia menanggung secara pribadi segala bentuk tuntutan hukum atas pelanggaran Hak Cipta dalam pembuatan Tugas Akhir ini.

Demikianlah pernyataan ini dibuat dan dapat dipergunakan sebagaimana mestinya.

Yogyakarta, 23 Juli 2025

Yang menyatakan,



Kendrick Chandra

200710554

HALAMAN PERSEMBAHAN

Semua akan indah pada waktu-Nya

KATA PENGANTAR

Puji dan syukur penulis haturkan kepada Tuhan Yang Maha Esa karena berkat rahmat dan karunia-Nya penulis dapat menyelesaikan pembuatan tugas akhir “Deteksi Malware pada File PDF dengan Algoritma SVM dan Random Forest” ini dengan baik.

Penulisan tugas akhir ini bertujuan untuk memenuhi salah satu syarat untuk mencapai derajat sarjana komputer dari Program Studi Informatika, Fakultas Teknologi Industri di Universitas Atma Jaya Yogyakarta.

Penulis menyadari bahwa dalam pembuatan tugas akhir ini penulis telah mendapatkan bantuan, bimbingan, dan dorongan dari banyak pihak. Untuk itu, pada kesempatan ini penulis ingin mengucapkan terima kasih kepada:

1. Tuhan Yang Maha Esa yang selalu membimbing, memberikan berkat-Nya, dan menyertai penulis selalu.
2. Dr. Ir. Parama Kartika Dewa SP., S.T., M.T., selaku Dekan Fakultas Teknologi Industri, Universitas Atma Jaya Yogyakarta.
3. Bapak Paulus Mudjihartono, S.T., M.T., Ph. D, selaku dosen pembimbing I yang telah membimbing dan memberikan masukan serta motivasi kepada penulis untuk menyelesaikan tugas akhir ini.
4. Ibu Vinindita Citrayasa, S.Pd., M.Hum, selaku dosen pembimbing II yang telah membimbing dan memberikan masukan serta motivasi kepada penulis untuk menyelesaikan tugas akhir ini.

Demikian laporan tugas akhir ini dibuat, dan penulis mengucapkan terima kasih kepada semua pihak. Semoga laporan ini dapat bermanfaat bagi pembaca.

Yogyakarta, 23 Juli 2025



Kendrick Chandra
200710554

DAFTAR ISI

LEMBAR PENGESAHAN	Error! Bookmark not defined.
PERNYATAAN ORISINALITAS & PUBLIKASI ILMIAH.....	iii
HALAMAN PERSEMBAHAN	iv
KATA PENGANTAR	v
DAFTAR ISI.....	vi
DAFTAR GAMBAR	viii
DAFTAR TABEL	ix
DAFTAR RUMUS	x
INTISARI.....	xi
BAB I	1
PENDAHULUAN	1
A. Latar Belakang	1
B. Rumusan Masalah	3
C. Batasan Masalah.....	3
D. Tujuan Penelitian	3
E. Metode Penelitian.....	5
F. Sistematika Penulisan	6
BAB II.....	8
TINJAUAN PUSTAKA	8
BAB III	14
LANDASAN TEORI.....	14
A. <i>File</i> PDF.....	14
B. <i>Malware</i> dalam <i>File</i> PDF.....	15
C. Analisis Statis.....	16
D. <i>Machine learning</i>	16
E. Support Vector Machine (SVM).....	18

F. Random Forest	20
G. Confusion Matrix	21
BAB IV	25
<i>DATASET</i> DAN PENGEMBANGAN MODEL.....	25
A. Deskripsi Problem.....	25
B. <i>Dataset</i>	25
1. Analisis Data	26
2. <i>Preprocessing</i> data	29
C. Pengembangan Model.....	35
BAB V.....	46
HASIL DAN PEMBAHASAN.....	46
A. Hasil Pengujian Model.....	46
B. Pembahasan Hasil Pengujian	53
BAB VI	55
PENUTUP.....	55
A. Kesimpulan	55
B. Saran.....	55
DAFTAR PUSTAKA	57

DAFTAR GAMBAR

Gambar 3.1 Gambar Hyperplane [25].....	18
Gambar 4.1. Operator Read CSV.....	29
Gambar 4.2. Operator Remove Duplicates	30
Gambar 4.3. Operator Nominal to Numerical.....	31
Gambar 4.4. Operator Set Role	32
Gambar 4.5. Operator Normalize.....	33
Gambar 4.6. Operator Split Data.....	34
Gambar 4.7. Diagram Alur Proses Pembangunan Model Deteksi Malware PDF .	36
Gambar 4.8. Operator Optimize Parameters (Grid).....	37
Gambar 4.9. Operator Cross Validation.....	39
Gambar 4.10. Operator Support Vector Machine (LibSVM)	41
Gambar 4.11. Operator Random Forest	42
Gambar 4.12. Operator Apply Model.	43
Gambar 4.13. Operator Performance	44
Gambar 5.1. AUC Average SVM	48
Gambar 5.2. AUC Optimis SVM.....	48
Gambar 5.3. AUC Pesimis SVM	49
Gambar 5.4. AUC Average Random Forest	51
Gambar 5.5. AUC Optimis Random Forest	52
Gambar 5.6. AUC Pesimis Random Forest	52

DAFTAR TABEL

Tabel 2.1 Tabel Perbandingan Literatur.....	11
Tabel 3.1 Confusion Matrix [27]	22
Tabel 5.1. Tabel Confusion Matrix SVM	46
Tabel 5.2. Confusion Matrix Random Forest	50
Tabel 5.3. Perbandingan SVM dan Random Forest.....	53

DAFTAR RUMUS

Rumus 3.1. Rumus Accuracy	23
Rumus 3.2. Rumus Precision	23
Rumus 3.3. Rumus Recall	23
Rumus 3.4. Rumus F1-Score	23
Rumus 5.1. Rumus Accuracy pada SVM	46
Rumus 5.2. Rumus Precision Kelas Malicious pada SVM	47
Rumus 5.3. Rumus Precision Kelas Benign pada SVM	47
Rumus 5.4. Rumus Recall Kelas Malicious pada SVM	47
Rumus 5.5. Rumus Recall Kelas Benign pada SVM	47
Rumus 5.6. Rumus F1-Score Kelas Malicious pada SVM	47
Rumus 5.7. Rumus F1-Score Kelas Benign pada SVM	47
Rumus 5.8. Rumus Accuracy pada Random Forest	50
Rumus 5.9. Rumus Precision Kelas Malicious pada Random Forest	50
Rumus 5.10. Rumus Precision Kelas Benign pada Random Forest	50
Rumus 5.11. Rumus Recall Kelas Malicious pada Random Forest	50
Rumus 5.12. Rumus Recall Kelas Benign pada Random Forest	50
Rumus 5.13. Rumus F1-Score Kelas Malicious pada Random Forest	51
Rumus 5.14. Rumus F1-Score Kelas Benign pada Random Forest	51

INTISARI

DETEKSI MALWARE PADA FILE PDF DENGAN ALGORITMA SVM DAN RANDOM FOREST

Intisari

Kendrick Chandra

200710554

Ancaman serangan digital yang terus meningkat di Indonesia, khususnya pada sektor pemerintahan sehingga menyoroti pentingnya sistem deteksi dini terhadap serangan digital, termasuk yang disisipkan melalui dokumen digital seperti *file* PDF. *File* PDF sering dimanfaatkan oleh pelaku kejahatan digital sebagai media penyebaran *malware* karena fleksibilitasnya dalam menyisipkan elemen-elemen aktif seperti JavaScript, *embedded files*, dan URI. Sayangnya, metode deteksi tradisional berbasis tanda tangan (*signature-based*) memiliki keterbatasan, terutama dalam mendeteksi jenis *malware* baru yang bersifat *polymorphic* dan *evasive*. Oleh karena itu, diperlukan pendekatan berbasis *machine learning* yang mampu mempelajari pola anomali secara adaptif. Penelitian ini difokuskan pada pengembangan dan evaluasi algoritma Support Vector Machine (SVM) untuk mendeteksi *malware* dalam *file* PDF, serta membandingkan kinerjanya dengan algoritma Random Forest.

Penelitian dilakukan dengan menggunakan *dataset* Evasive-PDFMal2022, yang terdiri dari 18.126 *file* PDF (9.063 *malicious* dan 9.063 *benign*). Tahapan penelitian dimulai dari studi literatur, pengumpulan dan analisis *dataset*, hingga *preprocessing* data menggunakan teknik normalisasi, *encoding* atribut kategorikal, serta pembagian data menjadi *training* dan *testing* set. Dua model klasifikasi kemudian dikembangkan, masing-masing menggunakan algoritma SVM dan Random Forest. Model dievaluasi menggunakan metrik *accuracy*, *precision*, *recall*, *F1-score*, dan AUC (Area Under Curve) untuk mengukur performa dalam membedakan *file benign* dan *malicious*. Proses implementasi dilakukan menggunakan RapidMiner dengan optimasi parameter melalui metode Grid Search dan Cross-Validation untuk meningkatkan akurasi dan menghindari *overfitting*.

Hasil evaluasi menunjukkan bahwa model SVM mampu mencapai akurasi sebesar 99,34% dengan nilai *F1-Score* yang seimbang antara kelas *benign* dan *malicious*. Sementara itu, model Random Forest menunjukkan performa lebih tinggi dengan *accuracy* 99,83% dan nilai *precision*, *recall*, serta *F1-score* yang

konsisten di atas 99,8% untuk kedua kelas. Perbandingan performa menunjukkan bahwa meskipun SVM cukup andal, Random Forest memiliki keunggulan dalam stabilitas dan akurasi deteksi. Penelitian ini menyimpulkan bahwa metode *machine learning*, khususnya SVM dan Random Forest, sangat potensial untuk diterapkan dalam sistem deteksi *malware* pada *file* PDF, dan dapat menjadi solusi alternatif terhadap keterbatasan metode deteksi konvensional.

Kata Kunci: PDF, *malware*, SVM, Random Forest, *machine learning*

Dosen Pembimbing I : Paulus Mudjihartono, S.T.,M.T., Ph. D.

Dosen Pembimbing II : Vinindita Citrayasa, S.Pd., M.Hum.

Jadwal Sidang Tugas Akhir : Kamis, 12 Januari 2025