BAB I PENDAHULUAN

1.1 Latar Belakang

Teknologi *deepfake*, yang didukung oleh algoritma kecerdasan buatan (*Artificial Intelligence*) berbasis pembelajaran mendalam (*deep learning*), memiliki kemampuan luar biasa untuk mengaburkan batas antara realitas dan manipulasi digital. Dengan memanfaatkan jaringan saraf tiruan, khususnya model seperti *Generative Adversarial Networks* (GANs), teknologi ini mampu mereplikasi karakteristik visual dan auditori, seperti fitur wajah dan intonasi suara, dengan tingkat presisi yang sangat tinggi, menghasilkan konten yang hampir tidak dapat dibedakan dari aslinya [1]. Namun, kemampuan ini menimbulkan tantangan signifikan dalam deteksi dan verifikasi keaslian konten, terutama bagi pengguna awam yang tidak memiliki akses ke alat analisis forensik canggih. Penyalahgunaan *deepfake* untuk tujuan merugikan, seperti penipuan daring, manipulasi opini publik, dan penyebaran disinformasi, semakin meningkat. Sebagai contoh, pada tahun 2023, berita di Indonesia menunjukkan kasus penyalahgunaan suara tokoh publik seperti Raffi Ahmad untuk mempromosikan perjudian daring dan suara palsu Najwa Shihab untuk mendukung kegiatan ilegal, yang mengecoh audiens dan merusak kepercayaan publik [2][3].

Pembuatan *deepfake* melibatkan analisis mendalam terhadap pola visual dan auditori dari data yang tersedia, seperti ribuan gambar wajah atau sampel suara, untuk membangun model digital yang realistis. Proses ini mencakup teknik *face mapping*, rekonstruksi tekstur, dan sinkronisasi gerakan, yang memungkinkan simulasi digital menyesuaikan ekspresi wajah, gerak tubuh, dan intonasi suara sesuai narasi yang diinginkan [1]. Pada tingkat lanjutan, teknologi ini dapat memodifikasi elemen-elemen tersebut secara dinamis, menciptakan manipulasi multimedia yang sangat sulit dibedakan dari kenyataan tanpa alat forensik. Kemampuan ini menjadikan *deepfake* alat yang kuat, namun juga rentan disalahgunakan untuk manipulasi identitas atau penyebaran disinformasi, sehingga memerlukan pendekatan teknis dan etis untuk mengelola dampaknya.

Deteksi *deepfake* menghadapi tantangan signifikan karena tingkat realisme yang tinggi dan kompleksitas manipulasi multimedia. Ketika konten *deepfake*

dipadukan dengan narasi yang tampak kredibel, kemampuan masyarakat umum untuk membedakan fakta dari fiksi semakin tergerus. Selain itu, implementasi dalam proses pengambilan keputusan model deteksi menjadi krusial, terutama dalam kasus hoaks yang melibatkan tokoh publik, seperti video manipulasi pejabat negara yang menyampaikan pernyataan palsu. Transparansi memungkinkan otoritas dan pengguna awam memahami fitur-fitur spesifik, seperti inkonsistensi gerakan wajah atau artefak audio, yang menjadi indikator manipulasi [7]. Kurangnya transparansi dapat melemahkan kepercayaan publik dan menghambat penerapan teknologi deteksi dalam konteks sosial atau hukum.

Untuk mengatasi tantangan deteksi *deepfake*, pendekatan berbasis *Deep learning* telah dikembangkan dengan mengintegrasikan *Convolutional Neural Networks* (CNN) dan *Capsule Networks* (CapsNet). CNN unggul dalam mengekstraksi fitur gambar, seperti tekstur wajah atau artefak manipulasi, melalui lapisan konvolusi yang mengidentifikasi pola kompleks [1]. Sebaliknya, CapsNet menawarkan pendekatan yang lebih robust dengan menangkap hubungan spasial dan hierarkis antar fitur melalui mekanisme *dynamic routing*, sehingga mampu mendeteksi ketidaksesuaian halus yang sering luput dari analisis konvensional [1]. Kombinasi ini memungkinkan analisis komprehensif terhadap konten multimedia, dengan CNN berfokus pada fitur tingkat rendah dan CapsNet menangkap struktur yang lebih kompleks, menghasilkan akurasi deteksi yang lebih tinggi, terutama pada video yang kompleks [1][4].

Dalam deteksi audio *deepfake*, CapsNet menunjukkan kemajuan signifikan melalui pendekatan perhatian berbasis kapsul (*capsule-based attention*). Setiap kapsul merepresentasikan aspek spesifik dari data audio, seperti karakteristik spektral atau temporal, dan berinteraksi untuk menghasilkan hierarki fitur yang kaya. Pendekatan ini memungkinkan mo Dalam deteksi audio deepfake, *Capsule Networks* (CapsNet) menunjukkan kemajuan signifikan melalui pendekatan perhatian berbasis kapsul (*capsule-based attention*), yang memungkinkan analisis data audio yang lebih robust dan terperinci. Setiap kapsul dalam arsitektur CapsNet merepresentasikan aspek spesifik dari data audio, seperti karakteristik spektral (*spectral rolloff, Mel-Frequency Cepstral Coefficients*), temporal (*zero-crossing*

rate, ritme), atau dinamika intonasi, dan berinteraksi melalui mekanisme *dynamic routing* untuk membentuk hierarki fitur yang kaya dan kontekstual [5]. Pendekatan ini memungkinkan model untuk secara selektif memfokuskan perhatian pada pola-pola penting yang mengindikasikan manipulasi, seperti anomali intonasi, artefak sintetik dari model *text-to-speech*, atau ketidaksesuaian transisi frekuensi, sehingga meningkatkan akurasi deteksi audio deepfake pada dataset yang bervariasi, termasuk data dengan noise atau variasi lingkungan [6].

Tantangan transparansi dalam deteksi deepfake diatasi melalui pendekatan *Explainable Artificial Intelligence* (XAI), yang mengatasi sifat *black-box* pada model seperti *Convolutional Neural Networks* (CNN) dan *Capsule Networks* (CapsNet). XAI memungkinkan model untuk menjelaskan alasan di balik klasifikasi dalam format yang mudah dipahami manusia, seperti visualisasi *heatmap* melalui teknik *Gradient-weighted Class Activation Mapping* (Grad-CAM) untuk menyoroti area manipulasi pada gambar (misalnya, ketidaksesuaian tekstur wajah) atau analisis fitur akustik seperti *Mel-Frequency Cepstral Coefficients* untuk mendeteksi anomali temporal pada audio [7]. Pendekatan ini mempermudah verifikasi oleh pengguna teknis dan non-teknis, sehingga meningkatkan kepercayaan publik terhadap keandalan sistem deteksi deepfake. Integrasi XAI dengan CNN dan CapsNet terbukti efektif dalam memberikan penjelasan yang jelas tentang fitur-fitur kunci, seperti distorsi tekstur, artefak visual, atau ketidaksesuaian intonasi, yang menjadi dasar deteksi manipulasi deepfake [7][8]. Dengan demikian, pendekatan ini mendukung penerapan teknologi yang lebih etis dan bertanggung jawab, terutama dalam konteks investigasi forensik digital dan verifikasi informasi publik, serta memitigasi risiko disinformasi dengan meningkatkan akuntabilitas dan transparansi sistem [8]. Lebih lanjut, penggunaan XAI seperti *SHAP* atau *LIME* dapat memperkaya interpretasi fitur-level, memungkinkan analisis yang lebih mendalam terhadap pola manipulasi yang kompleks, sehingga memperkuat kepercayaan dan penerapan praktis dalam skenario dunia nyata.

Penelitian ini mengusulkan pendekatan terpadu yang mengintegrasikan *Convolutional Neural Networks* (CNN), *Capsule Networks* (CapsNet), dan *Explainable Artificial Intelligence* (XAI) untuk mengembangkan sistem deteksi

deepfake multimodal yang akurat, transparan, dan andal. CNN dan CapsNet bekerja secara sinergis untuk mendeteksi manipulasi visual dan auditori dengan presisi tinggi, di mana CNN unggul dalam mengekstraksi fitur visual seperti tekstur wajah, gradien, dan artefak manipulasi, sementara CapsNet secara efektif menangkap hubungan spasial serta temporal antar fitur audio dan visual, seperti anomali intonasi atau ketidaksesuaian gerakan wajah [1]. Pendekatan XAI, melalui teknik seperti *Gradient-weighted Class Activation Mapping* (Grad-CAM) dan analisis fitur akustik (*Mel-Frequency Cepstral Coefficients* dan *spectral rolloff*), memastikan interpretabilitas hasil dengan memvisualisasikan area kritis pada gambar (misalnya, mata, hidung, mulut) dan anomali temporal pada audio, sehingga memungkinkan pengguna memahami dasar keputusan model [7].

1.2 Kajian Penelitian

Kajian penelitian adalah bentuk keaslian karya ilmiah yang akan dibuat sehingga memungkinkan untuk tidak adanya plagiat sebagai bentuk pencurian karya ilmiah. Kajian penelitian merupakan penjelasan perbandingan setiap penelitian yang telah dilakukan oleh peneliti sebelumnya dan menjadi referensi pembuatan tugas akhir penelitian. Tabel 1.1 merupakan cantuman referensi utama dengan penelitian terkait.

Tabel 1. 1 Penelitian Terkait.

No	Judul	Peneliti	Tahun
1	*Explainable Deepfake Video Detection using Convolutional Neural Network and CapsuleNet*	Gazi Hasin Ishrak, Zalish Mahmud, MD. Zami Al Zunaed Farabe, Tahera Khanom Tinni, Tanzim Reza, and Mohammad Zavid Parvez	2024
2	*ABC-CapsNet: Attention based Cascaded Capsule Network for Audio Deepfake Detection*	Taiba Majid Wani, Reeva Gulzar, and Irene Amerini	2024
3	*Detecting AI-generated images with CNN and*	Bharathi Mohan G, Prasanna Kumar Rangarajan, Akilesh rao S,	2024

No	Judul	Peneliti	Tahun
	Interpretation using Explainable AI	Mandava Sukesh, Abinandhini D M, and Jaikanth Y	
4	*DeepExplain: Enhancing Deepfake Detection Through Transparent and Explainable AI model*	Venkateswarlu Sunkari and A. Srinagesh	2024
5	*Capsule-Forensics Networks for Deepfake Detection*	Huy H. Nguyen, Junichi Yamagishi, and Isao Echizen	2022
6	*Detecting Deepfake Images Using Deep learning Techniques and Explainable AI Methods*	Wahidul Hasan Abir1, Faria Rahman Khanam, Kazi Nabiul Alam, Myriam Hadjouni, Hela Elmannai, Sami Bourouis, Rajesh Dey and Mohammad Monirujjaman Khan1	2022
7	*Explainability and Continuous Learning with Capsule Networks*	Janis Mohr1 , Basil Tousside1 , Marco Schmidt2 and Jorg Frochte	2021

Penelitian [1] dengan judul *Explainable Deepfake Video Detection using Convolutional Neural Network and CapsuleNet* menggabungkan *Convolutional Neural Network* (CNN) dengan *Capsule Networks* (CapsNet) untuk mendeteksi *deepfake* pada video. Dalam penelitian ini, CNN digunakan untuk mengekstraksi fitur visual dari video, sementara CapsNet digunakan untuk mempertahankan hubungan spasial antar fitur. Penelitian ini juga menekankan pentingnya *explainability* dengan memberikan interpretasi terhadap hasil deteksi. Namun, penelitian ini terbatas pada data visual, tanpa memperhatikan aspek audio yang sering digunakan dalam manipulasi.

Penelitian [5] dengan judul *ABC-CapsNet: Attention based Cascaded Capsule Network for Audio Deepfake Detection* memperkenalkan ABC-CapsNet,

yaitu *Attention-Based Cascaded Capsule Network*, untuk deteksi *deepfake* pada audio. Model ini memanfaatkan mekanisme perhatian berbasis kapsul yang memungkinkan fokus pada fitur-fitur penting dalam data audio, seperti intonasi dan pola frekuensi. Hasilnya menunjukkan peningkatan akurasi deteksi audio *deepfake*. Namun, penelitian ini belum mengintegrasikan aspek visual, yang merupakan bagian penting dalam *deepfake* multimodal.

Penelitian [7]. dengan judul *Detecting AI-generated images with* CNN *and Interpretation using Explainable* AI memadukan CNN dengan *Explainable* AI untuk mendeteksi gambar yang dihasilkan AI. Pendekatan ini memberikan transparansi dalam interpretasi hasil deteksi, tetapi tidak mengintegrasikan *Capsule Networks*, yang telah terbukti lebih unggul dalam representasi spasial.

Penelitian [11]. dengan judul *DeepExplain: Enhancing Deepfake Detection Through Transparent and Explainable* AI *Model* berfokus pada peningkatan deteksi *deepfake* melalui penerapan *Explainable* AI (XAI) yang transparan, memungkinkan model untuk tidak hanya mendeteksi manipulasi tetapi juga memberikan penjelasan yang dapat dimengerti manusia terkait hasil deteksi. Studi ini mengutamakan pengembangan model berbasis XAI untuk meningkatkan akurasi dan interpretabilitas dalam mendeteksi *deepfake*, terutama pada data visual, dengan tujuan menciptakan sistem yang dapat dipercaya oleh pengguna. Pendekatan ini relevan untuk menghadapi tantangan yang muncul dari penggunaan teknologi *deepfake* yang semakin canggih, dengan penekanan pada transparansi dalam proses deteksi.

Penelitian [4] dengan judul *Capsule-Forensics Networks for Deepfake Detection* memperkenalkan *Capsule-Forensics*, yang dirancang khusus untuk mendeteksi *deepfake*. Model ini menunjukkan kemampuan tinggi dalam mendeteksi manipulasi visual pada video, tetapi implementasi untuk data multimodal (audio dan visual) belum dibahas secara mendalam.

Penelitian [8] dengan judul *Detecting Deepfake Images Using Deep learning Techniques and Explainable* AI *Methods* berfokus pada pengembangan metode deteksi *deepfake* berbasis teknik *Deep learning* yang dikombinasikan dengan *Explainable* AI (XAI). Penelitian ini bertujuan untuk tidak hanya

meningkatkan akurasi deteksi manipulasi gambar *deepfake* tetapi juga memberikan interpretasi yang dapat dipahami oleh manusia mengenai bagaimana keputusan deteksi dibuat. Dengan memanfaatkan XAI, penelitian ini berusaha menciptakan model yang transparan dan dapat dipercaya, yang sangat penting dalam menghadapi ancaman teknologi *deepfake* yang semakin kompleks dan realistis.

Penelitian [12] dengan judul *Explainability and Continuous Learning with Capsule Networks* menyoroti pentingnya explainability dalam *Capsule Networks*, khususnya untuk aplikasi deteksi *deepfake*. Mekanisme explainability memungkinkan pengguna memahami bagaimana model mencapai keputusan deteksi, sehingga meningkatkan kepercayaan terhadap hasil. Namun, studi ini belum fokus pada penggabungan berbagai modalitas (audio dan visual) dalam deteksi *deepfake*.

Penelitian menawarkan pendekatan yang multimodal dalam deteksi *deepfake* dengan menggabungkan CNN, CapsNet, dan XAI, yang membedakannya dari penelitian lain yang hanya fokus pada satu aspek (audio atau visual). Berbeda dengan [1], yang hanya mengintegrasikan CNN dan CapsNet untuk video visual tanpa mempertimbangkan audio, penelitian menggabungkan kedua modalitas untuk mendeteksi *deepfake* secara lebih efektif. Selain itu, meskipun. [5] menekankan deteksi audio dengan model ABC-CapsNet, penelitian mereka tidak memasukkan aspek visual, sementara penelitian mengintegrasikan keduanya untuk mendeteksi *deepfake* multimodal. Penelitian. [7] juga menggabungkan CNN dengan XAI, tetapi tidak menggunakan CapsNet, yang lebih unggul dalam memahami hubungan spasial antar fitur. Fokus pada *interpretabilitas* dalam penelitian yang memanfaatkan XAI untuk implementasi keputusan deteksi, juga membedakan penelitian dari studi seperti [11], yang meskipun menggunakan XAI untuk visual, tidak mempertimbangkan CapsNet atau audio. Sementara itu. [4] memperkenalkan *Capsule-Forensics*, yang menggunakan CapsNet untuk mendeteksi manipulasi visual pada video, namun tidak mengintegrasikan audio, yang merupakan salah satu aspek penting dalam deteksi *deepfake* multimodal. [8] berfokus pada deteksi gambar *deepfake* dengan XAI, tetapi tidak mencakup aspek multimodal audio dan visual. Terakhir, [12] menyoroti pentingnya *explainability* dalam *Capsule*

Networks, namun tidak membahas integrasi multimodal untuk mendeteksi *deepfake* yang lebih kompleks. Dengan mengatasi keterbatasan penelitian sebelumnya, terutama dalam hal multimodalitas dan penggunaan CapsNet, penelitian memberikan solusi yang lebih efektif dan dapat dipercaya dalam mendeteksi manipulasi *deepfake* yang melibatkan audio dan visual, membuatnya lebih relevan untuk aplikasi dunia nyata.

1.3 Rumusan Masalah

Berdasarkan latar belakang yang telah dijelaskan, penelitian ini merumuskan permasalahan yang perlu diselesaikan, yaitu:

1. Bagaimana rancangan dan realisasi metode *Deep learning* dan *Explainable* AI untuk mendeteksi manipulasi gambar dan audio?
2. Bagaimana kinerja sistem deteksi manipulasi gambar dan audio menggunakan metode *Deep learning* dan *Explainable* AI?

1.4 Tujuan Penelitian

Berikut tujuan yang ingin dicapai dari penelitian ini adalah sebagai berikut:

1. Merancang dan mengimplementasikan metode *Deep learning* yang menggabungkan CNN, CapsNet, dan XAI untuk mendeteksi manipulasi pada gambar dan audio *deepfake*.
2. Menganalisis kinerja sistem deteksi manipulasi gambar dan audio yang menggunakan metode *Deep learning* dan *Explainable* AI

1.5 Manfaat Penelitian

Manfaat yang didapatkan serta diharapkan dari penelitian yang akan dilakukan adalah sebagai berikut:

1. Manfaat Akademik

   Penelitian ini memberikan kontribusi terhadap pengembangan ilmu pengetahuan di bidang kecerdasan buatan, khususnya pada deteksi *deepfake* berbasis citra wajah dan audio. Dengan mengintegrasikan algoritma terkini seperti CNN, CapsNet, dan pendekatan berbasis XAI, penelitian ini memperkaya literatur ilmiah serta menyediakan dasar teoritis yang kuat untuk penelitian lanjutan. Penelitian ini juga berkontribusi dalam pengembangan metode *deep learning*, yang menggabungkan keunggulan berbagai pendekatan

untuk meningkatkan akurasi dan efektivitas deteksi *deepfake*. Selain itu, penelitian ini mendorong eksplorasi lebih lanjut mengenai kombinasi teknik gambar dan audio untuk menciptakan solusi komprehensif dalam mendeteksi.

2. Manfaat Aplikatif

   Penelitian ini diharapkan menghasilkan sistem praktis yang mampu mendeteksi manipulasi gambar dan audio berbasis *deepfake*, khususnya dalam menangani penyalahgunaan wajah artis atau pejabat publik pada iklan ilegal, seperti promosi situs judi. Selain mendukung upaya pemerintah dan lembaga terkait dalam memerangi kejahatan berbasis teknologi ini, penelitian ini juga membantu meningkatkan kesadaran masyarakat terhadap ancaman *deepfake*. Lebih jauh, hasil penelitian ini dapat menjadi dasar bagi pengembangan kebijakan dan regulasi yang lebih efektif untuk memitigasi dampak negatif teknologi *deepfake*.

1.6 Batasan Masalah

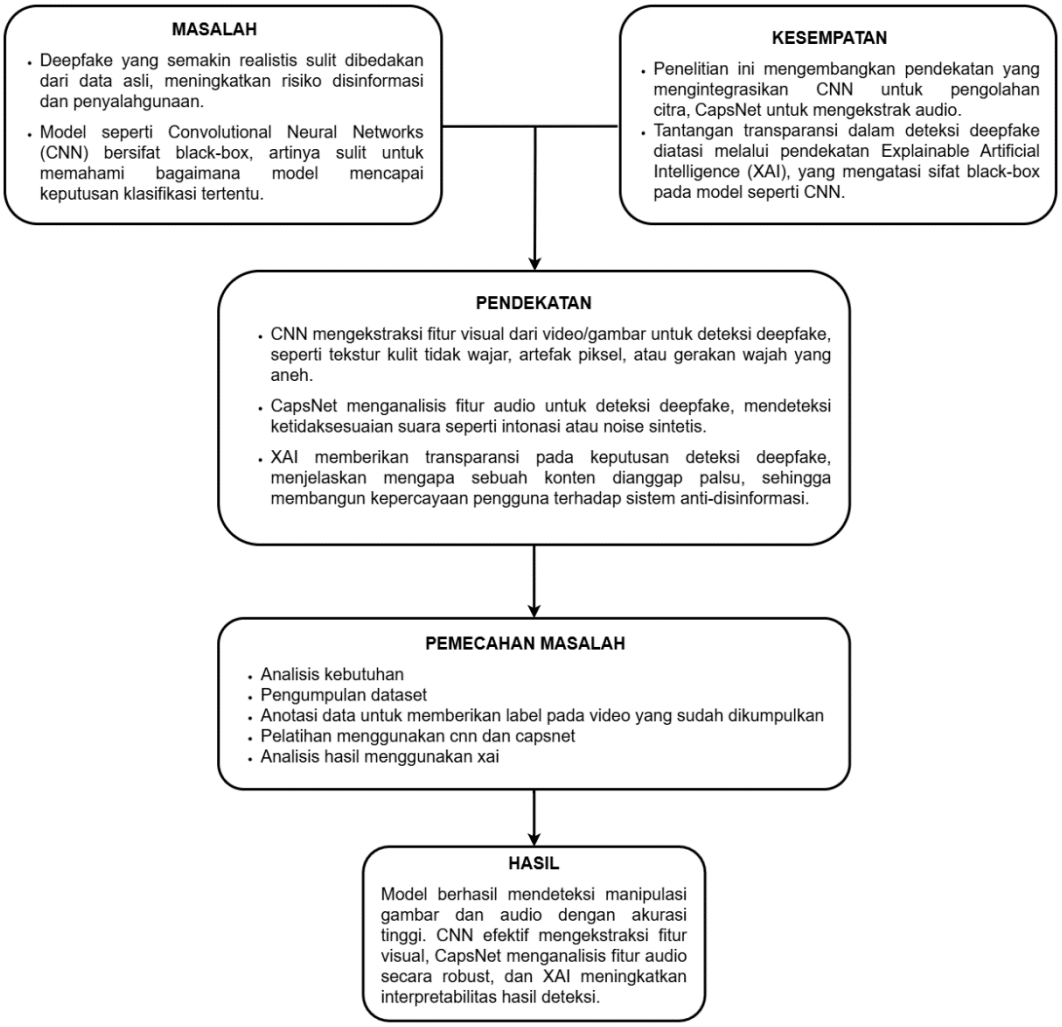
Batasan masalah dalam penelitian ini adalah sebagai berikut:

1. Data yang digunakan dalam penelitian ini terbatas pada gambar dalam format PNG dan audio dalam format WAV, sehingga berpotensi memengaruhi akurasi model dalam melakukan klasifikasi.
2. Pengujian model dilakukan langsung di lingkungan Google Colab dan tidak mencakup pengembangan antarmuka pengguna seperti aplikasi desktop, web, maupun mobile untuk implementasi atau pengujian lanjutan.
3. Dataset yang digunakan terbatas pada jenis manipulasi *deepfake* tertentu, seperti manipulasi wajah dan suara yang dihasilkan oleh algoritma tertentu. Variasi teknik manipulasi *deepfake* yang lebih baru atau kompleks, seperti *deepfake* berbasis diffusion models atau manipulasi parsial (splicing), mungkin tidak sepenuhnya tercakup, sehingga memengaruhi generalisasi model.

1.7 Kerangka Berpikir

Kerangka berpikir yaitu berisi alur pemikiran yang memuat uraian sistematis tentang hasil perumusan masalah penelitian yang diperkirakan dapat diselesaikan melalui pendekatan yang dibutuhkan untuk sistem deteksi *deepfake*.

Untuk mengatasi masalah tersebut, Kerangka berpikir penelitian ini dapat dijelaskan pada Gambar 1.1.



Gambar 1. 1 Kerangka Berpikir.

1.8 Sistematika Penulisan

Dalam mendapatkan struktur penyusunan data dan penulisan yang baik, laporan tugas akhir ini memiliki kerangka dan sistematika yang mengikuti aturan yangvtelah ditentukan, sehingga diharapkan mendapatkan hasil tulisan yang baik. Penulisan laporan tugas akhir ini mengikuti sistematika penulisan yang terdiri dari:

BAB I PENDAHULUAN

BAB I berisi gambaran umum dan dasar-dasar dalam penyususan skripsi sesuai dengan judul, seperti latar belakang, rumusan masalah, tujuan penelitian, manfaat hasil penelitian, batasan penelitian dan sistematika penulisan.

BAB II DASAR TEORI

BAB II berisi kajian kritis yang sistematis tentang aspek atau variable yang diteliti dengan menggunakan teori, konsep, dalil, ataupun peraturan yang relevan.

BAB III METODOLOGI PENELITIAN

BAB III berisi metodologi penelitian yang di dalamnya membahas tahapan tahapan yang diambil selama penelitian yang memuat jenis penelitian, sampel atau data, metode pengambilan data, jenis dan sumber data.

BAB IV PERANCANGAN DAN IMPLEMENTASI

BAB IV menjelaskan tentang penelitian yaitu dengan merancang model deteksi *deepfake* menggunakan *Convolutional Neural Network*s (CNN), *Capsule Networks* (CapsNet), dan *Explainable* AI (XAI).

BAB V HASIL DAN ANALISIS

BAB V berisi tentang hasil-hasil pengujian pada sistem yang telah dirancang. Pengujian sistem ini meliputi pengujian deteksi pada model dengan skenario video *real* dan *deepfake* yang sudah ditentukan. Kemudian melakukan pengujian terhadap data yang belum pernah di lihat oleh model.

BAB VI PENUTUP

BAB VI terdiri dari kesimpulan dari penelitian yang telah dilakukan, serta saran untuk penelitian-penelitian selanjutnya.