3. ANALISA DAN DESAIN SISTEM

3.1. Analisa Masalah

Pada penelitian sebelumnya tentang Facial Expression Recognition (FER), sebagian besar pendekatan hanya berfokus pada deteksi ekspresi wajah tunggal dalam satu frame atau citra. Implementasi algoritma FER umumnya menggunakan dataset yang memang dirancang khusus untuk deteksi wajah tunggal seperti CK+, JAFFE, dan FER-2013. Dataset-dataset tersebut biasanya memiliki satu wajah dominan dalam tiap gambar yang diberikan, tanpa mempertimbangkan skenario nyata dimana lebih dari satu wajah dapat muncul bersamaan dalam satu frame. Hal tersebut dapat menjadi masalah, apabila teknologi FER ini diterapkan secara nyata dalam berbagai aplikasi seperti robot pelayanan, sistem keamanan, atau interaksi sosial berbasis teknologi, kondisi nyata yang ditemui robot atau aplikasi tersebut kemungkinan besar tidak hanya terdiri dari satu wajah saja. Situasi ini menimbulkan tantangan tersendiri, karena model FER yang dilatih hanya menggunakan data dengan satu wajah mungkin tidak mampu mendeteksi ekspresi secara akurat pada situasi multi-wajah. Hal ini diperparah dengan ketersediaan dataset FER yang memuat gambar multi-person dengan berbagai emosi dalam satu frame masih sangat terbatas atau bahkan tidak tersedia sama sekali, sehingga menyebabkan perlunya pencarian atau pembuatan dataset baru yang secara khusus dirancang untuk kondisi multi-person dan multi-emotion guna meningkatkan representasi skenario nyata dalam dataset pelatihan.

Selain itu, sebagian besar dataset FER yang ada saat ini tidak dirancang untuk format pelatihan yang kompatibel dengan metode deteksi objek modern seperti YOLO (You Only Look Once). Model YOLO memerlukan format dataset berupa gambar yang dilengkapi dengan bounding box dan label untuk masing-masing objek yang akan dideteksi. Sedangkan dataset konvensional untuk FER umumnya hanya terdiri dari gambar wajah tunggal dengan anotasi ekspresi tanpa bounding box spesifik. Oleh karena itu, perlu dilakukan konversi dan penyesuaian dataset agar dapat digunakan secara efektif dengan algoritma YOLO, khususnya YOLO v11.

Selain masalah dataset dan kebutuhan untuk mendeteksi multi-person dan multi-emotion, pada penelitian ini juga terdapat masalah lain yaitu kebutuhan untuk mendeteksi wajah kecil (small faces). Deteksi wajah kecil sangat penting karena dalam situasi nyata, terutama pada sistem pengawasan, interaksi robot di ruang publik, atau video conference, wajah seseorang tidak selalu berada dekat dengan

kamera. Wajah kecil sering kali muncul akibat jarak yang jauh antara individu dengan kamera atau resolusi gambar yang rendah. Tanpa kemampuan mendeteksi wajah kecil, sistem FER berisiko gagal mengenali individu yang berada jauh atau tampak kecil dalam frame, sehingga menyebabkan penurunan akurasi deteksi emosi secara keseluruhan.

Penelitian ini bertujuan mengatasi tantangan tersebut dengan menyediakan pendekatan yang lebih adaptif terhadap situasi nyata, yaitu menggunakan dataset multi-wajah yang telah dianotasi sesuai format YOLO dan menerapkan algoritma deteksi ekspresi wajah berbasis YOLO v11 untuk meningkatkan akurasi dan kemampuan generalisasi model dalam situasi real-time dengan banyak wajah sekaligus, termasuk wajah berukuran kecil.

3.2. Analisa Kebutuhan

Dataset FER yang tersedia saat ini didominasi oleh gambar yang hanya berisi wajah tunggal, sehingga kurang representatif untuk menggambarkan kondisi nyata di mana sering kali terdapat lebih dari satu wajah dalam satu frame gambar. Oleh karena itu, diperlukan pendekatan baru dalam proses pelatihan YOLO v11 dengan menggunakan dataset yang berisi beberapa wajah dan emosi dalam satu gambar. Solusi yang diusulkan dalam penelitian ini adalah mengumpulkan dataset baru yang lebih representatif terhadap kondisi nyata, yaitu dataset yang terdiri dari gambar dengan satu atau beberapa orang yang memperlihatkan berbagai ekspresi wajah secara bersamaan. Dataset tersebut akan dikumpulkan melalui film, internet (seperti pencarian gambar melalui Google), dan sumber lain yang relevan, sehingga menghasilkan variasi ekspresi yang beragam dalam satu gambar.

Permasalahan lain dalam penelitian ini adalah tidak adanya format dataset FER yang kompatibel dengan model YOLO v11. Algoritma YOLO memerlukan format dataset dengan bounding box yang jelas dan label spesifik pada tiap objek wajah dalam gambar. Dataset konvensional yang tersedia saat ini biasanya hanya memberikan anotasi emosi tanpa informasi bounding box. Oleh karena itu, penelitian ini membutuhkan proses pelabelan dataset yang telah dikumpulkan sebelumnya. Proses ini mencakup pemberian bounding box secara manual pada tiap wajah yang muncul dalam gambar serta pemberian label yang menunjukkan ekspresi atau emosi yang ditampilkan oleh wajah tersebut. Dengan demikian, dataset hasil anotasi ini akan memenuhi persyaratan format YOLO v11, sehingga pelatihan dapat dilakukan secara efektif dan optimal.

Selain itu, solusi lain yang ditawarkan dalam penelitian ini untuk mengatasi permasalahan deteksi wajah kecil adalah dengan menerapkan teknologi Super-Resolution (SR). SR digunakan untuk meningkatkan kualitas dan resolusi gambar wajah kecil yang muncul dalam dataset, sehingga fitur wajah dapat dikenali lebih jelas oleh model deteksi. Penggunaan SR diharapkan mampu memperbaiki detail tekstur wajah kecil yang sebelumnya sulit dikenali oleh model, terutama dalam skenario nyata seperti pengawasan atau interaksi robot di ruang publik. SR akan diterapkan sebelum proses deteksi dilakukan, sehingga gambar wajah kecil akan direkonstruksi terlebih dahulu menjadi gambar dengan resolusi lebih tinggi, lalu dilanjutkan ke proses deteksi menggunakan YOLO v11. Pendekatan ini diharapkan dapat meningkatkan akurasi deteksi dan klasifikasi emosi pada wajah kecil dalam berbagai kondisi jarak dan kualitas gambar yang rendah.

3.3. Dataset

Dataset yang akan digunakan pada penelitian ini adalah sebuah dataset yang peneliti kumpulkan sendiri.

3.3.1. Pembuatan dataset

Dalam penelitian ini, peneliti membangun dataset secara mandiri karena tidak ditemukan dataset publik yang secara langsung sesuai dengan kebutuhan format dan struktur pelatihan model YOLO. Sebagian besar dataset *Facial Expression Recognition* (FER) yang tersedia hanya berisi satu wajah dengan posisi *close-up* di setiap gambar, tanpa adanya variasi jumlah wajah dalam satu frame. Padahal, model YOLO membutuhkan gambar dengan latar visual yang lebih kompleks, di mana satu gambar dapat memuat lebih dari satu wajah (multi-face) sekaligus mendeteksi ekspresi emosinya (multi-emotion). Struktur kelas emosi yang digunakan dalam dataset terdiri dari tujuh label utama yang mengacu pada standar Facial Expression Recognition, angry, disgust, fear, happy, neutral, sad, dan surprised.

Real Images diambil dari berbagai sumber terbuka seperti Google Images, poster film, dan media sosial. Gambar dari sumber ini cenderung natural dan tidak dibuat-buat. Namun, ditemukan kendala berupa ketimpangan distribusi emosi. Ekspresi yang sering muncul hanyalah happy, neutral, dan surprised, sementara ekspresi seperti fear, angry, dan disgust sulit ditemukan dalam jumlah cukup.

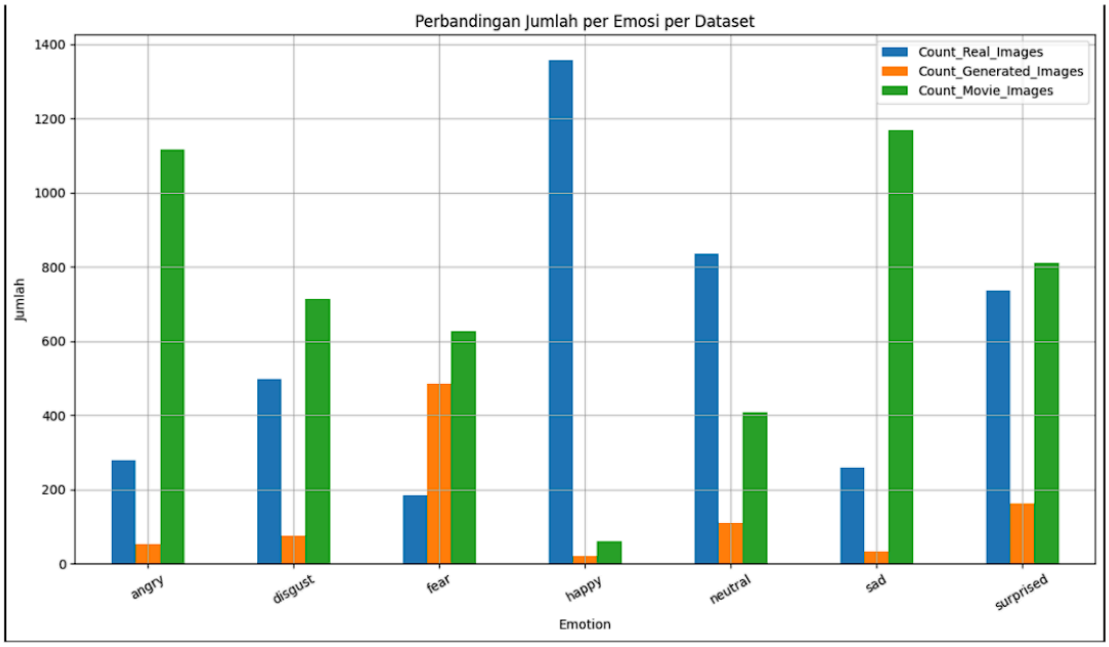
Untuk mengatasi permasalahan real images, peneliti mengekstraksi adegan dari film yang menampilkan karakter dengan ekspresi emosional intens. Sumber ini mampu menambahkan banyak data ekspresif terutama untuk emosi angry, sad, dan disgust. Meski begitu, ekspresi fear tetap menjadi salah satu kelas yang paling langka dan sulit didapat.

Untuk melengkapi distribusi dan mengatasi kekurangan gambar pada kelas fear, peneliti menggunakan teknologi generatif seperti text-to-image generation untuk menciptakan gambar sintetis. Hasilnya dari gambar generatif digunakan agar proporsi jumlah gambar per emosi menjadi lebih seimbang.

Tabel 3.1 Jumlah Gambar per Emosi dan Sumber Dataset

Emotion	**Real**	**Movie**	**Generated**	**Total**
angry	278	1117	52	**1447**
disgust	497	713	74	**1284**
fear	185	626	486	**1297**
happy	1358	60	20	**1438**
neutral	835	408	109	**1352**
sad	259	1169	34	**1462**
surprised	735	811	161	**1707**
	4147	**4904**	**936**	**9987**

Gambar 3.1 Perbandingan Jumlah Gambar per Emosi dan Sumber Dataset

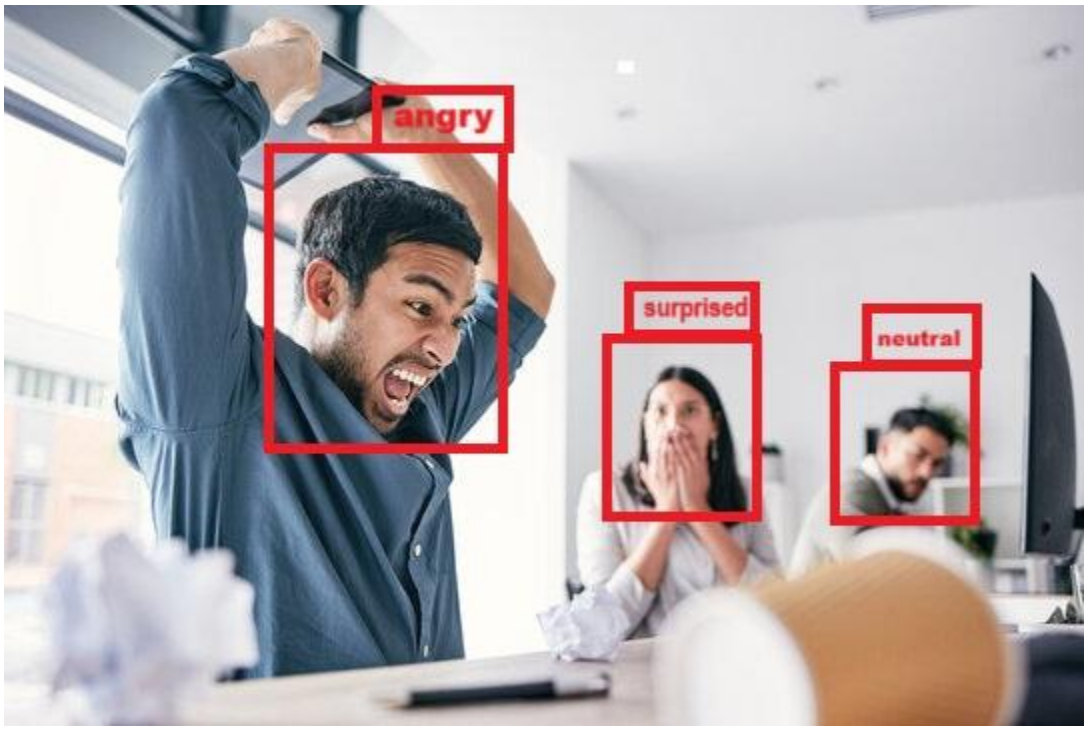


Gambar 3.1 contoh gambar yang pada dataset yang akan dibuat penulis

3.3.2. Merubah Format dataset menjadi Format train YOLO

Model YOLO v11 ini memiliki format train sendiri yang berbeda dengan model-model yang lain. Format tersebut adalah model YOLO meminta untuk labelingnya dilakukan beserta dengan bounding

box. Jadi untuk mentrain pada model ini diperlukan format dataset yang sudah pada gambarnya terdapat bounding box dan label tiap bounding box tersebut. Untuk tugas ini nantinya pada penelitian ini akan dilakukan dengan menggunakan roboflow. Alat tersebut nantinya akan dapat membantu untuk membuat bounding box dan label untuk dataset yang peneliti kumpulkan.



Gambar 3.2 contoh dataset yang telah diberi label dan bounding box

3.4. Desain Sistem

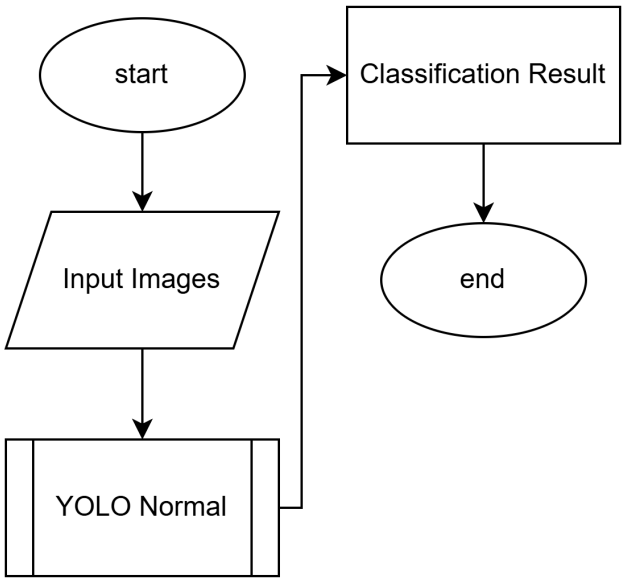
Desain sistem pada penelitian ini disusun untuk menangani permasalahan pengenalan ekspresi wajah (Facial Expression Recognition/FER) dalam kondisi multi-wajah dan multi-emosi secara real-time. Sistem dirancang untuk menguji dan membandingkan empat arsitektur utama yang digunakan dalam penelitian, model YOLOv11 Normal, YOLOv5, YOLOv11 + ESRGAN (Enhanced Super-Resolution GAN), Face Detection Model + YOLOv11-CLS (Classifier).

3.4.1. YOLOv11 dan YOLOv5 (End-to-End Detection and Classification)

Pendekatan pertama dalam penelitian ini adalah dengan menggunakan model deteksi dan klasifikasi ekspresi secara end-to-end, yaitu model YOLOv11 dan YOLOv5. Kedua model menerima gambar input berisi beberapa wajah, lalu secara langsung menghasilkan output berupa bounding box dan label emosi pada masing-masing wajah.

Model YOLOv11 merupakan versi terbaru dari keluarga YOLO yang dilengkapi dengan backbone baru, attention mechanism, dan optimalisasi struktur, yang membuatnya unggul dalam mendeteksi objek kecil dan kompleks seperti ekspresi wajah. Sedangkan YOLOv5, meskipun merupakan versi terdahulu, digunakan sebagai pembanding performa karena telah banyak digunakan dalam berbagai studi FER. Kedua model ini diuji menggunakan dataset yang sama dan konfigurasi hyperparameter identik agar hasil pengujian dapat dibandingkan secara objektif. Tujuan pengujian ini adalah untuk menilai sejauh mana perbedaan arsitektur mempengaruhi akurasi klasifikasi ekspresi wajah.

Gambar 3.3 *Flowchart* alur deteksi *Facial Expression Recognition*



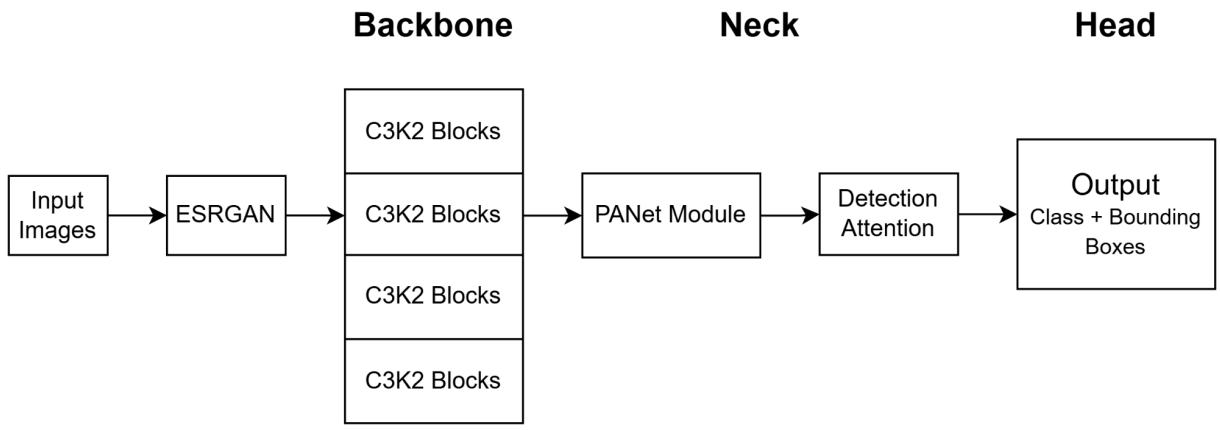
3.4.2. YOLOv11 + ESRGAN (Super-Resolution)

Metode kedua adalah integrasi antara model YOLOv11 dan teknologi super-resolusi ESRGAN (Enhanced Super-Resolution Generative Adversarial Network). Pendekatan ini dirancang untuk meningkatkan performa deteksi ekspresi terutama pada wajah kecil (small faces) yang sering muncul dalam gambar real-world seperti foto grup atau cuplikan video.

Alur sistem dalam pendekatan ini terdiri dari dua tahap utama. Pertama adalah tahap preprocessing, gambar input diproses oleh ESRGAN untuk meningkatkan resolusi, khususnya pada bagian wajah yang kecil dan buram. Lalu selanjutnya adalah tahap deteksi dan klasifikasi, dimana gambar hasil peningkatan resolusi kemudian diproses oleh model YOLOv11 untuk mendeteksi wajah serta mengklasifikasikan ekspresinya. Dengan rekonstruksi visual dari wajah kecil, fitur-fitur halus dapat

diperjelas sehingga proses ekstraksi fitur oleh YOLO menjadi lebih optimal. Metode ini diharapkan mampu mengatasi kelemahan deteksi ekspresi pada wajah berukuran kecil yang umum terjadi dalam kondisi nyata.

Gambar 3.4 Ilustrasi model YOLO v11 + SR dalam sistem deteksi *multi-person and multi-emotion*



3.4.3. Face Detection + YOLOv11-CLS (Pipeline Dua Tahap)

Metode ketiga menggunakan pendekatan dua tahap, yang memisahkan proses deteksi wajah dan klasifikasi ekspresi ke dalam dua model terpisah. Alur sistemnya sebagai berikut:

1. Deteksi Wajah: Gambar input diproses oleh model YOLOv11-M yang telah dilatih khusus untuk mendeteksi wajah saja (label: “Face”). Setiap wajah dalam gambar akan diberikan bounding box oleh model ini.
2. Pemangkasan (Cropping): Setiap wajah yang terdeteksi kemudian di-crop secara otomatis berdasarkan koordinat bounding box.
3. Klasifikasi Ekspresi: Gambar wajah hasil crop dikirim ke model YOLOv11-CLS, yaitu model klasifikasi yang hanya menerima input wajah dan mengoutput salah satu dari 7 label emosi (Angry, Disgust, Fear, Happy, Neutral, Sad, dan Surprised).

Pendekatan ini bertujuan untuk mengurangi noise latar belakang dan meningkatkan akurasi klasifikasi, karena model klasifikasi hanya fokus pada area wajah. Pendekatan ini juga memungkinkan spesialisasi model: satu model dilatih hanya untuk mendeteksi wajah, dan model lain dilatih hanya untuk klasifikasi emosi.

Gambar 3.5 Ilustrasi model Face Detection + YOLOv11-CLS