MoodMe Benchmarks: Jan 2024

By Sebastian Algharaballi-Yanow (ML ENG & Co-Founder)

Benchmark Metrics

Accuracy

Percentage of faces the model is correctly predicting on the entire set of data.

Number of correct predictions

÷

Total number of predictions

Precision

Percentage of images correctly classified within a specific category over the total number of images **classified** as belonging to that category. Essentially measures **False Positives.**

True Positives

÷

True Positives + False Positives

Recall

Percentage of images correctly classified within a specific category over the total number of images **actually belonging** to that category. Essentially measures **False Negatives.**

True Positives

÷

True Positives + False Negatives

The F1 Score is the harmonic average between precision and recall. **Support** is the number of faces/files detected in a particular category. **Undetected faces** is the percentage of faces in the test datasets that a model was unable to read and classify.

1

Emotion

1.1

Emotion Models

Models Used For Emotion Analysis

- MoodMe
  - Our reference model.
- UMONS
  - Direct competitor and partner of MoodMe.
- OpenVINO
  - Very well-known set of deep learning models optimized for Intel software.
- Deepface w/OpenCV Backend
  - Known for efficiency and fast processing speeds. One of the most popular models. Developed by AI engineers at Facebook.
- Deepface w/Retinaface Backend
  - Known for superior performance in detected faces under “real world” photo conditions. A more “in-depth” version of OpenCV. Also developed by AI engineers at Facebook.

General Notes on Models:

- Each test was conducted with preprocessed 48x48 images being fed into the models. This ensures fair comparison across all tests.
- Since OpenVINO can only process 5 emotions compared to the standard 7, “fear” and “disgust” are omitted from the accuracy calculation. However, “fear” and “disgust” will still be analyzed for precision and recall in our categorical metrics.
- Each model **strictly runs on the edge**, making it fair to compare to MoodMe.

1.2

RAFDB

Dataset Description

Real-world Affective Faces Database (RAF-DB)

DETAILS

- ~ 30k diverse facial images downloaded from the internet
- Consists of a single-label subset, including 7 classes of basic emotions and two-tab subset, including 12 classes of compound emotions. The “basic” subset was used for this experiment (15399 files)
- Emotions: surprise, fear, disgust, happy, sad, angry, neutral
- Realistically reflects multiple demographics and real-world photo scenarios. Chosen because it proves a realistic and challenging benchmark for the models to be tested on.
- Also used for GAE prediction tasks.

LINKS

- Details
- Research Paper
- Dataset folder

1.3

Emotion Benchmarking Results: RAFDB

Model Benchmarks (RAF)

Model Used for Test	Emotion Accuracy	Undetected (Emotion)	Runtime (Emotion)
MoodMe	26%	0%	42 seconds
UMONS	18%	0%	44 seconds
OpenVINO	17%	0%	17 seconds
OpenCV (with Deepface)	25%	84.68%	1 minute
Retinaface (with Deepface)	21%	50.62%	30 minutes

Notes:

- MoodMe was the best model for this test in terms of accuracy, detected faces, and overall runtime.
- The RAF dataset houses many complex images, some of which vary in orientations, backgrounds, and overall quality of the photos. Therefore, obtaining an accuracy of 26% is a good accomplishment, especially in such a short amount of time.

Emotion Categorical Benchmarks (RAF)

	precision	recall	f1-score	support
Anger	0.20	0.06	0.09	1619
Disgust	0.05	0.08	0.06	355
Fear	0.00	0.00	0.00	877
Happiness	0.41	0.41	0.41	5957
Sadness	0.15	0.23	0.18	2460
Surprise	0.07	0.27	0.12	866
Neutral	0.21	0.08	0.12	3204

MoodMe:

- Performs best at detecting happiness and struggles the most with fear. However, the dataset is heavily skewed towards categories like happy, sad, and neutral, and has a limited amount of the instances that the model struggles with. This can be a potential confounder. More data could lead to better results.
- Holds the best overall categorical results when compared the other models.

	precision	recall	f1-score	support
Anger	0.12	0.03	0.05	1619
Disgust	0.02	0.06	0.04	355
Fear	0.06	0.04	0.05	877
Happiness	0.39	0.07	0.11	5957
Sadness	0.13	0.18	0.15	2460
Surprise	0.05	0.05	0.05	866
Neutral	0.20	0.53	0.30	3204

UMONS:

- Performs best when presented when neutral faces, but poor performance everywhere else. Once again, it is not a coincidence that the model performs the best when presented with the three majority categories of the dataset (happy, sad, and neutral).

Emotion Categorical Benchmarks (RAF)

	precision	recall	f1-score	support
Surprise	0.23	0.11	0.15	207
Fear	0.03	0.26	0.05	53
Disgust	0.10	0.37	0.16	145
Happiness	0.65	0.31	0.42	1001
Sadness	0.19	0.31	0.23	277
Anger	0.07	0.01	0.02	111
Neutral	0.44	0.20	0.27	556

OpenCV:

- Small support due to lack of ability to process a substantial amount of faces. Results are based on ~15% of the total dataset

	precision	recall	f1-score	support
Surprise	0.35	0.12	0.18	940
Fear	0.01	0.26	0.02	86
Disgust	0.08	0.32	0.13	475
Happiness	0.70	0.27	0.39	3113
Sadness	0.19	0.29	0.23	1069
Anger	0.13	0.03	0.05	329
Neutral	0.33	0.14	0.20	1562

Retinaface:

- Multiple categories with higher precision than recall. This means that the model is very selective when predicting positive instances, leading to many positive instances being overlooked.
- Results are based on ~50% of the dataset due to undetected faces

1.4

FER

Dataset Description

Facial Expression Recognition: FER-2013

DETAILS

- ~ 7k 48x48 grayscale images. Faces have been automatically registered so that the face is more or less centered and occupies about the same amount of space in each image.
- Emotions: surprise, fear, disgust, happy, sad, angry, neutral
- Possesses a concise and comprehensive set of aligned, lab-like photos in a controlled environment. Chosen to test each model's core emotion recognition abilities with one face and one expression per photo.

LINKS

- FER
- Research Paper

1.5

Emotion Benchmarking Results: FER

Model Benchmarks (FER)

Model Used for Test	Emotion Accuracy	Undetected (Emotion)	Runtime (Emotion)
MoodMe	70%	0%	20 seconds
UMONS	43%	0%	20 seconds
OpenVINO	67%	0%	7 seconds
OpenCV (with Deepface)	58%	0%	2 minutes
Retinaface (with Deepface)	57%	0%	52 minutes

Notes:

- MoodMe was again the best model for this test in terms of accuracy, detected faces, and overall runtime.
- Much higher accuracy metrics across the board with this dataset, with MoodMe being the highest. The model does an excellent job at processing the small, black and white images in this dataset.

Emotion Categorical Benchmarks (FER)

	precision	recall	f1-score	support
anger	0.56	0.64	0.60	958
disgust	0.78	0.61	0.69	111
fear	0.55	0.45	0.49	1024
happy	0.87	0.85	0.86	1774
sadness	0.60	0.46	0.52	1247
surprise	0.73	0.83	0.78	831
neutral	0.57	0.70	0.63	1233

MoodMe:

- Excellent marks across the board. Model shows flexibility in processing small faces as a result of being trained on similar photos (of course, benchmarks were made on the test set)
- Best overall results once again

	precision	recall	f1-score	support
anger	0.07	0.06	0.06	958
fear	0.36	0.02	0.05	1024
disgust	0.04	0.06	0.05	111
happy	0.66	0.78	0.71	1774
sadness	0.37	0.22	0.28	1247
surprise	0.06	0.02	0.03	831
neutral	0.29	0.73	0.41	1233

UMONS:

- Poor performance on multiple categories (besides happiness). Falsely predicts a large proportion of faces as neutral (precision metric).

Emotion Categorical Benchmarks (FER)

	precision	recall	f1-score	support
angry	0.50	0.08	0.13	958
neutral	0.57	0.25	0.35	1233
happy	0.90	0.57	0.70	1774
sad	0.40	0.27	0.32	1247
surprise	0.52	0.66	0.58	831
fear	0.27	0.38	0.31	1024
disgust	0.05	0.94	0.10	111

OpenCV:

- Multiple instances of high precision with low recall and vise versa. Model had a difficult time balancing false positives and false negatives.

	precision	recall	f1-score	support
angry	0.55	0.04	0.07	4790
neutral	0.64	0.15	0.25	6165
happy	0.88	0.32	0.47	7096
sad	0.33	0.18	0.24	3741
surprise	0.48	0.45	0.46	2493
fear	0.23	0.26	0.24	3072
disgust	0.04	0.63	0.07	333

Retinaface:

- Much bigger support than the other models due to precise facial recognition of the model. This is interesting since each picture in the dataset houses only one face, but these results indicate the model recognized multiple faces in multiple pictures.
- Similar results to the model above. Many instances of high precision and low recall, and vise versa.

1.6

CK+

Dataset description

Extended Cohn-Kanade: CK+

DETAILS

- ~ 1k 48x48 photos (2304 pixels) gathered from ~600 video sequences from 123 different subjects. Widely regarded as one of the most extensively used lab-controlled facial expression classification datasets available.
- Emotions: surprise, fear, disgust, happy, sad, angry, neutral, and contempt (which was not included in this test)
- Similar to FER, was chosen as a test dataset due to lab-controlled setup. Adds an extra dimension of realistic analysis being that the photos are taken from video sequences.

LINKS

- CK+
- Research Papers

1.7

Emotion Benchmarking Results: CK+

Model Benchmarks (CK+)

Model	Accuracy	Undetected Faces	Runtime
MoodMe	84%	0%	2 seconds
UMONS	44%	0%	4 seconds
OpenVINO	79%	0%	3 seconds
OpenCV (with Deepface)	68%	0%	15 seconds
Retinaface (with Deepface)	66%	0%	6 minutes

Notes:

- MoodMe was the best model by far when analyzing this dataset. Achieving over 80% accuracy in less than 3 seconds is remarkable.
- UMONS continues to struggle, while the three other models continue to perform at around the same level.

Emotion Categorical Benchmarks (CK+)

	precision	recall	f1-score	support
anger	0.47	0.74	0.57	135
disgust	1.00	0.47	0.64	177
fear	0.66	0.44	0.53	75
happy	0.98	1.00	0.99	207
sadness	0.51	0.46	0.48	84
surprise	0.93	0.88	0.90	249
neutral	0.40	0.81	0.54	54

MoodMe:

- Good, balanced benchmarks. Excellent precision marks for “disgust”, “happy” and “surprise”. Relatively high false positive rates for “neutral”, “anger”, and “sadness”
- Best overall results

	precision	recall	f1-score	support
anger	0.00	0.00	0.00	135
disgust	0.00	0.00	0.00	177
fear	0.00	0.00	0.00	75
happy	0.88	1.00	0.94	207
sadness	0.73	0.77	0.75	84
surprise	0.06	0.00	0.01	249
neutral	0.14	0.94	0.24	54

UMONS:

- Very poor overall performance, especially on emotions other than “happy” and “sad”
- Model tends to falsely predict an image as neutral (indicated by the low precision value)

Emotion Categorical Benchmarks (CK+)

	precision	recall	f1-score	support
angry	0.43	0.02	0.04	135
disgust	0.41	0.98	0.58	177
fear	0.24	0.29	0.26	75
happy	0.98	0.86	0.92	207
sad	0.28	0.08	0.13	84
surprise	0.89	0.70	0.79	249
neutral	0.46	0.48	0.47	54

OpenCV:

- Lots of misses (false negatives) on “anger” and “sadness”, and many false positives on “disgust”. Decent results overall, but does not compete with MoodMe

	precision	recall	f1-score	support
angry	0.43	0.02	0.04	135
disgust	0.42	0.97	0.59	177
fear	0.19	0.31	0.23	75
happy	0.97	0.85	0.91	207
sad	0.32	0.13	0.19	84
surprise	0.91	0.61	0.73	249
neutral	0.49	0.54	0.51	54

Retinaface:

- Very similar results as above.

2

Gender

2.1

Gender Models

Models Used For Gender Analysis

- MoodMe/UMONS
- Deepface w/OpenCV Backend
- Deepface w/Retinaface Backend
- Insightface
    - Very popular gender/age model that was released under an MIT license. Specializes in recognizing faces through different poses and occlusion.

General Notes on Models:

- Instead of 48x48 image processing, the MoodMe/UMONS Gender model processes 96x96 images. Therefore, all images fed into other models were set to 96x96 to ensure fair comparison.
- OpenVINO does not have a gender model, so we will use InsightFace as a replacement.
- All models are capable of computing GAE (besides Insightface, which only calculates gender and age).
- As before, all models run **on the edge.**

2.2

Gender Benchmarking Results: RAFDB

Gender Model Benchmarks (RAF)

Model Used for Test	Gender Accuracy	Undetected (Gender)	Runtime (Gender)
MoodMe/UMONS	44.13%	0%	15 seconds
OpenCV/Deepface	37.15%	84.68%	2 minutes
Retinaface/Deepface	52.67%	50.62%	33 minutes
Insightface	31.80%	0%	1.5 minutes

Notes:

- MoodMe/UMONS had the best overall performance when accounting a balance between processing speed and accuracy.
- While Retinaface higher accuracy, the model was not able to recognize faces in over 50% of the dataset, making their results non-significant. It also took ~33 minutes for Retinaface to complete processing.

- NOTE: The MoodMe/UMONS model was trained on the RAF dataset. The model's exposure to the RAF dataset during training may influence its performance on the testing set. Therefore, these results will be omitted from our final summary to avoid potential bias. However, they are still displayed here to validate the model's performance in comparison to others.

Gender Categorical Benchmarks (RAF)

	precision	recall	f1-score	support
Male	0.4224	0.9341	0.5818	6206
Female	0.6619	0.1307	0.2183	8181

MoodMe/UMONS:

- Many false positives for males, but excellent at not missing any male predictions.
- Poorer performance recognizing females. Often misses pictures that should be classified as female.
- Still has the best overall performance.

	precision	recall	f1-score	support
Male	0.3520	0.9975	0.5203	802
Female	0.9481	0.0479	0.0911	1525

OpenCV/Deepface:

- Small support due to lack of ability to detect faces. Moderate performance on male faces, but very poor performance on female faces (*many* false negatives).

Gender Categorical Benchmarks (RAF)

	precision	recall	f1-score	support
Male	0.4361	0.9668	0.6011	3130
Female	0.7748	0.1233	0.2128	3989

Retinaface/Deepface:

- High miss rate for female faces, and a high false positive rate for male faces.
- Similar performance to the MoodMe model.

	precision	recall	f1-score	support
Male	0.22	0.22	0.22	5971
Female	0.39	0.42	0.41	7615
Unsure	0.00	0.00	0.00	730

Insightface:

- Nearly even distribution between precision and recall for the respective male and female categories.
- Metrics are simply not as high as the other models.

2.3

UTKFace

Dataset description

UTKFace

DETAILS

- ~ 24k “in the wild” face images with labels for age, gender, and ethnicity. Images contain real-world variations in pose, facial expressions, illumination, and resolution.
- Similar to RAF, this dataset will provide an excellent standard for face recognition in natural states. Over 20,000 images is also a great dataset size to draw conclusions from.

LINKS

- UTKFace
- Research Papers

2.4

Gender Benchmarking Results: UTKFace

Gender Model Benchmarks (UTKFace)

Model	Accuracy	Undetected	Runtime
MoodMe/UMONS	65.68%	0%	1 minute
OpenCV/Deepface	85.73%	48.95%	10 minutes
Retinaface/Deepface	74.33%	2.54%	1 hour
Insightface	9.52%	1.09%	3 minutes

Notes:

- Retinaface/Deepface showcases the best overall performance. However, it did take the model an hour to achieve these results, while MoodMe wasn't that far behind after a processing time of 1 minute.
- The MoodMe results are very promising due to the short computation time compared to Retinaface, and one could argue that the tradeoff between a slightly less accurate model that is 60x faster than a more accurate model is worth it.

Gender Categorical Benchmarks (UTKFace)

	precision	recall	f1-score	support
Male	0.62	0.86	0.72	12581
Female	0.74	0.43	0.55	11523

MoodMe/UMONS:

- Good overall performance. Detected males at a higher rate than females.
- Showed a high false negative (miss) rate for females, and a higher false positive rate for males.

	precision	recall	f1-score	support
Male	0.8075	0.9666	0.8799	6652
Female	0.9487	0.7289	0.8244	5654

OpenCV:

- Very high metrics across the board. Would be the best model (by far) if it was able to recognize more faces in the dataset (~50% undetected rate).

Gender Categorical Benchmarks (UTKFace)

	precision	recall	f1-score	support
Male	0.6816	0.9524	0.7945	12242
Female	0.9087	0.5159	0.6582	11250

Retinaface/Deepface:

- Best overall performance after looking at the detailed metrics.
- Higher false negative rate for females compared to males, but still shows excellent all-around performance.

	precision	recall	f1-score	support
Male	0.1024	0.0938	0.0979	12479
Female	0.0886	0.0967	0.0925	11363

Insightface:

- *Very* poor marks across the board. Model nearly failed to complete the task at hand.

2.3

FairFace

Dataset description

FairFace

DETAILS

- Known as one of the most well-balanced and least biased datasets used for classifying gender, age, and ethnicity.
- Dataset consists of ~90k training set images and ~11k validation set images. Tests were completed on the smaller validation set that mainly consists of “in-the-wild” photos.
- Chosen due to size and overall balance between categories. Will be easier to draw correlations from a balanced dataset.

LINKS

- FairFace
- Research Paper

2.4

Gender Benchmarking Results: FairFace

Gender Model Benchmarks (FairFace)

Model	Accuracy	Undetected	Runtime
MoodMe	61.41%	0%	13 seconds
OpenCV (with Deepface)	74.49%	58.77%	4 minutes
Retinaface (with Deepface)	71.37%	1%	1.5 hours
InsightFace	17.88%	0%	1 minute

Notes:

- MoodMe/UMONS shows the best overall performance. While the model didn't achieve the highest accuracy compared to the others, a runtime of only 13 seconds cannot be ignored.
- Retinaface achieved 10% better accuracy, but took 1.5 hours to do so, which is not very practical in real-world model deployment.

Gender Categorical Benchmarks (UTKFace)

	precision	recall	f1-score	support
Male	0.59	0.90	0.71	5792
Female	0.72	0.30	0.42	5162

MoodMe/UMONS:

- Better at detecting males compared to females.
- Higher false positive rate for males, and very high false negative (miss) rate for females.

	precision	recall	f1-score	support
Male	0.6640	0.9546	0.7832	2180
Female	0.9284	0.5492	0.6902	2336

OpenCV:

- Higher false positive rate for males, and higher false negative rate for females.
- Similar to MoodMe, this means the model misses a lot of actual female instances, and captures most of the male instances but also incorrectly labels some instances as male when they are not.
- Lower support due to many undetected faces.

Gender Categorical Benchmarks (UTKFace)

	precision	recall	f1-score	support
Male	0.6611	0.9406	0.7765	5787
Female	0.8732	0.4591	0.6018	5158

Retinaface:

- Falls into the same predicament as the other models do: missing a lot of actual female instances, and capturing most of the male instances but also incorrectly labeling instances as male when they are female.
- Produces the best pure scores out of the group, but as stated before, takes over an hour to achieve this.

	precision	recall	f1-score	support
Male	0.1740	0.1477	0.1598	5790
Female	0.1827	0.2137	0.1970	5161

Insightface:

- Poor performance across the board.

3

Age

3.1

Age Benchmarking Results: RAFDB

Age Model Benchmarks (RAF)

Model	Age Accuracy	Undetected (Age)	Runtime (Age)
MoodMe/UMONS	36%	0%	7 minutes
OpenCV/Deepface	68%	84.68%	1.5 minutes
Retinaface/Deepface	53%	50.62%	33 minutes
Insightface	4%	0%	1.5 minutes

- NOTE: The MoodMe/UMONS model was trained on the RAF dataset. The model's exposure to the RAF dataset during training may influence its performance on the testing set. Therefore, these results will be omitted from our final summary to avoid potential bias. However, they are still displayed here to validate the model's performance in comparison to others.

Notes:

- Each image was preprocessed to size 224x224 before testing due to the requirements of the MoodMe/UMONS model.
- While MoodMe didn't showcase the highest accuracy, it held the best overall performance when considering runtime and detected faces.

Age Categorical Benchmarks (RAF)

	precision	recall	f1-score	support
0-3	0.00	0.00	0.00	1612
4-19	0.36	0.01	0.02	2657
20-39	0.53	0.49	0.51	8192
40-69	0.19	0.59	0.28	2422
70+	0.06	0.00	0.00	455

MoodMe/UMONS:

- Best at categorizing ages 20-39 (majority class of the dataset). Many false negatives on ages 4-19, and many false positives on ages 40-69.
- Still holds the best overall performance.

	precision	recall	f1-score	support
0-3	0.00	0.00	0.00	51
4-19	1.00	0.00	0.01	207
20-39	0.69	0.99	0.81	1618
40-69	0.21	0.02	0.03	432
70+	0.00	0.00	0.00	42

OpenCV/Deepface:

- Small support pool due to inability of recognizing faces. Best results between ages 20-39 (majority class). *High* miss rate for ages 4-19 and 40-69.

Age Categorical Benchmarks (RAF)

	precision	recall	f1-score	support
0-3	0.00	0.00	0.00	830
4-19	0.71	0.01	0.01	1374
20-39	0.53	0.99	0.69	3999
40-69	0.22	0.02	0.04	1212
70+	0.00	0.00	0.00	159

Retinaface/Deepface:

- Similar performance to OpenCV, except now with a larger support.
- Massive inability to detect ages 4-19 and 40-69.

	precision	recall	f1-score	support
0-3	0.27	0.00	0.00	10870
4-19	0.09	0.01	0.02	1081
20-39	0.17	0.23	0.20	2365
40-69	0.00	0.00	0.00	0
70+	0.00	0.00	0.00	0

Insightface:

- Very low metrics across the board. Majority of faces were incorrectly classified as ages 0-3. Failed to compute ages older than 40.

3.2

Age Benchmarking Results: UTKFace

Age Model Benchmarks (UTKFace)

Model	Accuracy	Undetected	Undetected
MoodMe/UMONS	56%	0%	13 minutes
OpenCV/Deepface	58%	24.71%	9 minutes
Retinaface/Deepface	57%	2.93%	52 minutes
Insightface	62.11%	1.10%	3 minutes

Notes:

- Insightface achieved the best overall performance.
- One problem with the MoodMe/UMONS age model is that it is currently trained to process images of size 224x224. This leads to longer processing times and potentially less accurate results.

Age Categorical Benchmarks (UTKFace)

	precision	recall	f1-score	support
0–3	0.97	0.07	0.13	2121
4–19	0.48	0.24	0.32	2768
20–39	0.69	0.59	0.64	11911
40–69	0.43	0.83	0.56	5892
70+	0.69	0.45	0.54	1414

MoodMe/UMONS:

- Very large miss-rate (false negatives) for ages 0-3. Good performance in the other categories, even in non-majority classes.

	precision	recall	f1-score	support
0–3	0.00	0.00	0.00	1185
4–19	0.39	0.06	0.10	1956
20–39	0.58	0.97	0.73	9379
40–69	0.63	0.28	0.38	4747
70+	0.00	0.00	0.00	882

OpenCV:

- Poor job detecting elders and babies. High miss rate for ages 4-19. High model accuracy comes from great performance on ages 20-39.
- Smaller support due to lack of ability to recognize faces.

Age Categorical Benchmarks (UTKFace)

	precision	recall	f1-score	support
0–3	0.00	0.00	0.00	2087
4–19	0.43	0.02	0.04	2714
20–39	0.57	0.95	0.71	11604
40–69	0.57	0.39	0.46	5689
70+	0.00	0.00	0.00	1306

Retinaface:

- Similar results as OpenCV, except with a larger support.
- Overall accuracy is, once again, misleading due to the model's high performance on ages 20-39 with low performance everywhere else.

	precision	recall	f1-score	support
0–3	0.9680	0.3128	0.4728	2030
4–19	0.1806	0.0726	0.1036	2700
20–39	0.7232	0.7469	0.7349	11863
40–69	0.5282	0.7098	0.6057	5864
70+	0.4843	0.6885	0.5686	1387

Insightface:

- Best performance out of all the models (low computation time, highest overall accuracy, and most balanced precision and recall scores)
- Highest precision/recall on ages 20-39 (majority class). Many false negatives (misses) on ages 0-3.

3.2

Age Benchmarking Results: FairFace

Age Model Benchmarks (Fairface)

Model	Accuracy	Undetected	Runtime
MoodMe	31%	0%	5 minutes
OpenCV	29%	34.6%	7 minutes
Retinaface	29%	1%	1.5 hours
Insightface	28%	0%	3 minutes

Notes:

- Low overall performance across all models, but MoodMe still performed the best.
- Similar patterns as other tests; MoodMe/UMONS detects faces in all photos at a quicker speed than other models.

Age Categorical Benchmarks (FairFace)

	precision	recall	f1-score	support
0-2	0.00	0.00	0.00	199
3-9	0.00	0.00	0.00	1356
10-19	0.23	0.14	0.18	1181
20-29	0.40	0.35	0.38	3300
30-39	0.28	0.43	0.34	2330
40-49	0.25	0.43	0.32	1353
50-59	0.31	0.39	0.35	796
60-69	0.32	0.39	0.35	321
more than 70	0.36	0.24	0.29	118

MoodMe/UMONS:

- Difficult time classifying infants and young children. Strongest at classifying adults aged 20-29, but still shows low results across the board.
- Albeit not the best results one would hope for, the model outperformed the rest of its competitors.

	precision	recall	f1-score	support
0-2	0.00	0.00	0.00	120
3-9	0.00	0.00	0.00	886
10-19	0.07	0.01	0.01	781
20-29	0.35	0.43	0.38	2198
30-39	0.25	0.60	0.36	1542
40-49	0.28	0.21	0.24	871
50-59	0.40	0.04	0.07	479
60-69	0.33	0.00	0.01	204
more than 70	0.00	0.00	0.00	81

OpenCV:

- Poor job detecting elders, babies, young children, and teens. Stronger performance when classifying adults aged 20-39.
- Lower support due to lack of ability to recognize faces in all contexts.

Age Categorical Benchmarks (FairFace)

	precision	recall	f1-score	support
0-2	0.00	0.00	0.00	199
3-9	0.00	0.00	0.00	1356
10-19	0.08	0.00	0.01	1181
20-29	0.34	0.33	0.33	3298
30-39	0.27	0.72	0.39	2328
40-49	0.28	0.28	0.28	1352
50-59	0.36	0.06	0.11	796
60-69	0.25	0.01	0.01	321
more than 70	0.00	0.00	0.00	118

Retinaface:

- Like OpenCV, poor performance on children, teens, and the elderly.
- High false positive rate on ages 30-39. Most balanced results on ages 20-29, but with low metrics.

	precision	recall	f1-score	support
0-2	0.65	0.09	0.15	199
3-9	0.49	0.06	0.11	1356
10-19	0.06	0.01	0.01	1181
20-29	0.35	0.36	0.36	3299
30-39	0.28	0.44	0.34	2330
40-49	0.23	0.31	0.27	1353
50-59	0.22	0.26	0.24	796
60-69	0.17	0.26	0.21	321
more than 70	0.21	0.41	0.28	118

Insightface:

- Very high false negative rate on infants, meaning the model has a hard time detecting infants when presented. Same can be said about children.
- Like other models, showcases the best performance from ages 20-29, which is the majority class of the dataset.

4

Ethnicity

4.1

Ethnicity Benchmarking Results: RAFDB

Ethnicity Model Benchmarks (RAF)

Model	Accuracy	Undetected	Runtime
MoodMe/UMONS	61.77%	0%	20 seconds
OpenCV/Deepface	78.50%	84.68%	2 minutes
Retinaface/Deepface	73.36%	50.62%	33 minutes

- NOTE: The MoodMe/UMONS model was trained on the RAF dataset. The model's exposure to the RAF dataset during training may influence its performance on the testing set. Therefore, these results will be omitted from our final summary to avoid potential bias. However, they are still displayed here to validate the model's performance in comparison to others.

Notes:

- MoodMe/UMONS shows the best overall performance. Model does an excellent job at detecting all faces with fast processing speeds and a high accuracy.
- Higher accuracy of other models do not hold as much weight when they can't process over half the dataset.
- Insightface is omitted from these results since the model can compute only gender and age.

Ethnicity Categorical Benchmarks (RAF)

Note: RAFDB only has available ground-truth labels for White, Black, and Asian faces.
Therefore, the categorical benchmarks will only be measuring these ethnicities.

	precision	recall	f1-score	support
White	0.7732	0.7562	0.7646	10404
Black	0.1070	0.0276	0.0439	1050
Asian	0.1686	0.2398	0.1979	2173

MoodMe/UMONS:

- Excellent in categorizing white (majority class) faces.
- Struggles with Asian and Black faces, but on a smaller sample size.
- Best overall performance

	precision	recall	f1-score	support
White	0.88	0.89	0.88	1937
Black	0.44	0.16	0.23	89
Asian	0.28	0.30	0.29	295

OpenCV/Deepface:

- Small sample size due to lack of ability to recognize all faces in the dataset.
- Good performance on white faces, and decent performance on Asian faces.
- Lacks face recognition power of MoodMe/UMONS model.

Ethnicity Categorical Benchmarks (RAF)

	precision	recall	f1-score	support
White	0.82	0.86	0.84	5715
Black	0.50	0.24	0.32	533
Asian	0.34	0.33	0.34	1207

Retinaface/Deepface:

- Would claim the best overall results if it had the ability to recognize more faces in the dataset (undetected ~50%).
- Many misses for Black faces, with even distribution in precision/recall for White and Asian faces (majority sets).

4.2

Ethnicity Benchmarking Results: UTKFace

Ethnicity Model Benchmarks (UTKFace)

Model	Accuracy	Undetected	Runtime
MoodMe/UMONS	30%	0%	1 minute
OpenCV/Deepface	69.15%	48.95%	10 minutes
Retinaface/Deepface	63.05%	2.55%	1 hour

Notes:

- Retinaface outperformed the other models, even with ~1 hour processing time.

Ethnicity Categorical Benchmarks (UTKFace)

Note: UTKFace has ground truth metrics available for White, Black, Asian, and Indian, and “Other” faces, which includes Hispanic, Latino, and Middle Eastern.

	precision	recall	f1-score	support
White	0.54	0.47	0.50	10222
Black	0.17	0.06	0.09	4558
Asian	0.23	0.41	0.29	3586
Indian	0.39	0.06	0.10	4027
Others	0.06	0.24	0.10	1710

MoodMe/UMONS:

- Performs well on the majority class of white faces, but struggles with the rest of the classes.
- High miss rate for Indian faces, and a higher false positive rate for Asian faces. Poor overall performance on Black faces.

	precision	recall	f1-score	support
White	0.84	0.82	0.83	5526
Black	0.88	0.71	0.79	1716
Asian	0.68	0.82	0.74	1555
Indian	0.77	0.43	0.55	2499
Others	0.18	0.40	0.25	1010

OpenCV/Deepface:

- Much better performance compared to MoodMe/UMONS, even with a lower support due to weak face recognition.
- Very balanced overall results. Higher miss rate for Indian faces. Would be claimed as the best model if it could recognize more faces (undetected rate of ~50%)

Ethnicity Categorical Benchmarks (UTKFace)

	precision	recall	f1-score	support
White	0.76	0.78	0.77	9912
Black	0.87	0.65	0.74	4416
Asian	0.53	0.79	0.63	3497
Indian	0.71	0.24	0.36	3984
Others	0.15	0.30	0.20	1682

Retinaface/Deepface:

- Slightly decreased performance compared to OpenCV/Deepface, but is still claimed as the best due to higher facial recognition capabilities.
- Mostly balanced, with the exceptions of a high miss rate for Indian faces, and a higher false negative rate for Asian faces.

4.2

Ethnicity Benchmarking Results: FairFace

Ethnicity Model Benchmarks (FairFace)

Model	Accuracy	Undetected	Runtime
MoodMe/UMONS	23.81%	0%	15 seconds
OpenCV	64.17%	56.45%	8 minutes
Retinaface	56.16%	1%	1.5 hours

Notes:

- OpenCV and Retinaface showed better performance than MoodMe/UMONS.
- While MoodMe/UMONS was still the fastest and most reliable model when it came to actually detecting faces, ~24% accuracy is not close to what the other two models were able to achieve.

Ethnicity Categorical Benchmarks (FairFace)

	precision	recall	f1-score	support
East Asian	0.22	0.42	0.29	1550
Southeast Asian	1.00	0.37	0.54	1415
Indian	0.28	0.06	0.10	1516
Latino_Hispanic	0.13	0.24	0.17	1623
White	0.30	0.35	0.32	2085
Middle Eastern	0.16	0.12	0.14	1209
Black	0.07	0.05	0.06	1556

MoodMe/UMONS:

- Low metrics across the board, including a high miss rate for Southeast Asian and Indian faces.

	precision	recall	f1-score	support
East Asian	0.71	0.94	0.81	687
Southeast Asian	1.00	0.83	0.91	648
Indian	0.68	0.50	0.58	707
Latino_Hispanic	0.47	0.38	0.42	865
White	0.59	0.70	0.64	927
Middle Eastern	0.44	0.43	0.43	538
Black	0.64	0.77	0.70	398

OpenCV/Deepface:

- Much better metrics than MoodMe/UMONS, albeit with a lower support.
- Excellent performance on all Asian faces, with moderate performance everywhere else.

Ethnicity Categorical Benchmarks (FairFace)

	precision	recall	f1-score	support
East Asian	0.52	0.87	0.65	1547
Southeast Asian	1.00	0.78	0.88	1415
Indian	0.65	0.27	0.39	1515
Latino_Hispanic	0.36	0.25	0.30	1623
White	0.47	0.67	0.56	2084
Middle Eastern	0.42	0.29	0.35	1207
Black	0.66	0.72	0.69	1554

Retinaface/Deepface:

- Slightly lower results compared to OpenCV, but expected due to the larger support.
- Showcases the best performance on Black faces, an ethnicity that the two other models struggled with.

Executive Summary

Emotion Detection: MoodMe excelled across multiple datasets, and was consistently the best performer when compared to rival models. The processing times were a highlight and something that should be emphasized when marketing (along with being the most accurate model out of all edge emotion AI models tested). Outperformed the average accuracy of our competitors by ~30% while demonstrating remarkable speed (~20x faster than the average competitor).

Emotion Improvements: Recognizing less common emotions, such as “fear” and “neutral”, is something that the model can be tuned to fix.

Gender Detection: The MoodMe/UMONS model excels at processing faces faster than any other competitor. The model was also arguably the best performer across all test datasets when taking both accuracy and computation time into consideration.

Gender Improvements: The model shows a tendency to generate more false positives for male faces and false negatives for female faces, meaning that, at times, it struggles to identify women.

Age Detection: Model does well at classifying individuals aged 20-39.

Age Improvements: The model that needs the most improvement. Slower processing times due to being programmed to process only 224x224 images. Struggles with classifying infants, children, and the elderly.

Ethnicity Detection: Processes faces faster than all other ethnicity competitors. Shows promising results when classifying Caucasian and Asian faces.

Ethnicity Improvements: Struggles mightily with overall accuracy and distinguishing between multiple ethnic groups. Was heavily outperformed (accuracy wise) by rival models in all significant tests.