Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X
IMPACT FACTOR: 7.056

IJCSMC, Vol. 15, Issue. 6, June 2026, pg.52 – 60

Can Facial Micro-Expressions Reveal Lies? An Experiment Using Machine Learning

Zehra Grbovic[1]; Zerina Altoka[2]

[1]Department of Information Technology, International Burch University, Bosnia and Herzegovina
[2]Department of Information Technology, International Burch University, Bosnia and Herzegovina
[1] zehra.grbovic@stu.ibu.edu.ba; [2] zerina.masetic@ibu.edu.ba

DOI: https://doi.org/10.47760/ijcsmc.2026.v15i06.005

Abstract: This paper examines whether brief, involuntary facial movements known as micro-expressions can support automatic discrimination between deceptive and truthful statements. Micro-expressions last less than a quarter of a second and are hard to control on purpose. We used a public dataset of 320 short videos (80 people, 4 videos each) where each person told two truths and two lies. We measured facial muscle movements in every video frame using OpenFace, and used these measurements to try to predict lying versus telling the truth. We made sure to test our model on people it had never seen before, which is a stricter and more honest test than many earlier studies use. Our best model correctly identified lies and truths 58.8% of the time, significantly above the 50% chance level (p=0.001). We also built a separate, simpler method that only looked for very short facial flickers matching the exact definition of a micro-expression. This method identified a genuine statistical pattern: one specific eye movement (Action Unit 05, the upper-lid raiser) occurred more frequently during deceptive statements. However, classification using only this pattern performed no better than chance. We explain why this happened. The few clues we found were too closely related to each other to give the computer enough separate information to learn from. Our main conclusion is that detecting micro-expressions on their own can confirm a real difference between liars and truth-tellers, but is not yet enough, by itself, to reliably catch lies. Combining it with broader measures of facial movement produced a numerically higher accuracy, though this improvement over the broader measures alone was not found to be statistically significant.

Keywords: lie detection, micro-expressions, facial action units, OpenFace, machine learning, random forest, person-independent testing

I. INTRODUCTION

People are not very good at spotting lies. On average, humans guess correctly only about 54% of the time [1]. This has led researchers to investigate whether automated systems can outperform human judgment by detecting subtle physiological signals that are not consciously perceptible.

One such signal is the micro-expression. A micro-expression is a very brief facial movement, lasting only about 0.04 to 0.5 seconds, that is thought to happen when someone is hiding their true feelings [2], [3], [4], [5]. Because it happens so fast, a person usually cannot stop it from happening, even if they are trying hard to keep a straight face. This makes micro-expressions a theoretically compelling target for automated deception detection.

In the past, researchers had to watch videos frame by frame and mark these tiny movements by hand. This took a long time and different people might disagree on what they saw [2]. Now, software tools such as OpenFace [6], [7] can measure facial muscle movements automatically, frame by frame, across an entire video. OpenFace recognizes Facial Action Units (AUs) [6], which are numbered terms used to describe facial movement [8]. For example, AU05 denotes the upper eyelid is raised, and AU12 means the corner of the mouth is pulled up, as in a smile.

This paper uses OpenFace measurements to test whether micro-expressions can help tell lies apart from truths. The focus is on doing this test in a fair and honest way. In related areas of behavior and physiology research, it is a well-known risk that testing a model on the same people it was trained on can make results look better than they really are, since the computer can partly learn to recognize a specific person instead of learning the general behavior being studied [9]. In our study, every person used for testing was completely new to the model, the model had never seen that person's face during training.

This paper makes the following contributions:

- Test of a machine learning model for lie detection using a fair, person-independent method, where no person appears in both the training data and the test data.
- Design of the facial movement measurements that are specifically built to catch brief, sudden movements, instead of just averaging movement over a whole video, which would hide short bursts.
- A separate, rule-based detector that looks only for movements matching the exact timing of a micro-expression (very short, sudden bursts), and test of whether this narrower, more targeted approach can predict lying on its own.
- Report of the negative result: the targeted micro-expression detector finds a real difference between liars and truth-tellers, but this difference is not strong enough by itself to make good predictions.

II. RELATED WORK

A. *Why Micro-Expressions Might Reveal Lies*

The basic idea behind micro-expressions comes from Ekman [2], who proposed that real emotions “leak out” through the face even when a person is trying to hide them, and that this leakage is especially likely to show up as a brief, involuntary movement, because such movements happen too fast to consciously control. Some studies have found that trained observers can use micro-expressions to detect lies better than chance [10]. Other studies have found weaker results [11], so this remains a topic with mixed evidence.

Certain face movements have been linked to lying in past research. AU14 (a dimple near the mouth corner) has been linked to a feeling of contempt [12]. AU23 and AU24 (lightening the lips and pressing the lips) can appear during emotional control or deception-related situations involving increased cognitive load [13]. AU05 (upper eyelid raise) and AU02 (outer brow raise) are components of the facial surprise and fear expressions, involving eye-widening and brow-raising patterns [14], [15]. AU26 (jaw drop) is likewise a component of the surprise expression, typically co-occurring with AU01, AU02, and AU05 [15].

B. *Computer-Based Lie Detection*

Before automatic tools existed, researchers had to mark facial movements by hand, which was slow and not always consistent between different researchers [2]. The OpenFace tool [6] changed this by measuring AUs automatically from video, about as accurately as a trained human observer, but much faster and for as many videos as needed.

A few studies have already tried building lie detectors using facial movements, sometimes together with other clues such as voice or word choice. Wu et al. [16] combined facial movement features with audio and text features and reported a strong result (0.877 area under the curve) on a dataset of real courtroom trial videos. Perez-Rosas et al. [17] also combined facial movement with verbal and other non-verbal cues, reporting accuracies between 60% and 75% on a similar real-trial dataset. Both of these results came from combining facial movement with other modalities, rather than using facial movement alone, which leaves an open question about how well facial movement by itself can perform, and how well it can perform when tested fairly on completely new people. This paper focuses specifically on that open question.

A small number of studies have tested whether micro-expressions specifically can be used to automatically detect lies, and the results have generally been disappointing. Jordan et al. found that training people to recognize micro-expressions did not meaningfully improve their ability to detect lies [11]. Dinges et al. [18]

point out that micro-expressions are a weak signal for this task for a structural reason: they happen during both honest and deceptive moments, and they occur too rarely and too inconsistently from person to person to reliably tell the two apart. In their own experiments combining multiple micro-expression datasets to get enough training data, they found that even though their model learned something useful when tested on the same dataset it was trained on, it performed no better than a random guess when tested on a different dataset. Based on this, they recommend against relying on micro-expressions for deception detection until future datasets can address these limitations [18]. The present paper revisits this open question directly, asking whether micro-expressions, used on their own, can support reliable lie detection.

C. The Dataset Used in This Study

Miami University Deception Detection Database (MU3D) [19] is a public dataset built specifically for lie detection research. It includes 80 people, each recorded saying four short statements: a positive truth statement, a negative lie statement, a negative truth statement, and a positive lie statement. "Positive" and "negative" simply refer to whether the statement was about something the person liked or disliked. This paper does not focus on that difference, we only care whether each statement was a lie or the truth. The dataset has an equal number of men and women and people of different racial backgrounds.

Although previous machine-learning studies have demonstrated that facial behavior contributes useful information for deception detection, the highest-performing systems typically combine facial, vocal, and linguistic features. This raises an important question regarding the predictive value of facial micro-expressions alone. The present study addresses this question by evaluating whether deception can be detected using only micro-expression features extracted from the MU3D dataset.

III. DATASET AND DATA CLEANING

MU3D dataset has 320 videos in total (80 people, each with 4 videos). Videos were labeled as "Truth" and "Lie," giving us exactly 160 truth videos and 160 lie videos, a perfectly even split, which makes later testing simpler and fairer. We ran OpenFace on every video, which measured 17 different face muscle movements. Videos were about 30-40 seconds long, recorded at 30 frames per second, giving us roughly 1,000 frames of data per video.

Not every video frame was good enough to use. OpenFace gives a confidence score for how sure it is about its face-tracking in each frame. Any frame where this confidence was below 70%, or where OpenFace flagged the tracking as failed, was removed. If a few frames in a row were missing after this step, they were filled in by smoothly estimating values between the frames before and after the gap, but only for small gaps (5 frames or fewer). Any clearly broken values were fixed, like an AU score outside its normal range. After all this cleaning, almost no data was lost (295 of 320 videos kept every single frame, and even the worst affected video only lost about 7% of its frames).

To make sure that the results were trustworthy, we used a testing method called Leave-One-Subject-Out, or LOSO [20]. We picked one person, set aside all four of their videos as the "test," and trained the model using everyone else's videos. We then check if the model can correctly guess lie versus truth for the person it has never seen. This process was repeated 80 times, once for every person in the dataset, so that everyone gets a turn being the test case. This way, the model is never tested on someone it already learned from, which gives us an honest measure of how well it would work on a new person in the real world.

IV. MEASURING FACIAL MOVEMENT

A micro-expression lasts less than 200 milliseconds, which is only about 1 to 6 video frames at 30 frames per second. If we simply average a face movement score over an entire 1,000-frame video, a brief 3-frame spike gets averaged together with 997 frames of almost no movement. The spike becomes practically invisible in the average. This means we needed a different way to measure movement that could still notice these brief spikes, instead of smoothing them away.

For each of the 17 face muscle measurements, several different numbers were calculated that describe how that muscle moved throughout the video, focusing on shape and sudden change rather than simple averages:

- Distributional shape (kurtosis and skewness): these measures indicate whether the movement signal remained near zero with occasional sharp spikes, the pattern expected from a micro-expression.
- How quickly the movement changed from one frame to the next: a sudden jump is more likely to be a real micro-expression than a slow, gradual change.
- The number of sudden intensity spikes occurring in the video.
- How long each burst of movement lasted: micro-expressions are defined by being short, so we counted how many bursts lasted 6 frames or fewer.

We also added three combined measurements based on known facial expression patterns: a "genuine smile" score (cheek and mouth corner moving together), a "tension" score (brow and eyelid tightening together), and a "fake smile" score (mouth corner moving without the cheek, which can indicate a forced smile). For each of the

17 AU intensity channels, 13 statistical measurements were computed: four distributional shape measurements (maximum, standard deviation, skewness, kurtosis), three frame-to-frame difference measurements (maximum, standard deviation, and mean of frame-to-frame change), two spike measurements (spike count and spike rate), and four peak-duration measurements (mean peak duration, minimum peak duration, peak count, and the micro-burst ratio). This yields 221 measurements (17 × 13). An additional 18 measurements captured the activation rate of each binary AU indicator, and the remaining 6 measurements were the three combined expression scores (genuine smile, tension, and fake smile), each computed as both a mean and a maximum value. This produced 245 measurements per video in total (221 + 18 + 6 = 245).

With 245 numbers but only 320 videos, there is a real risk that the model could "memorize" quirks of the training data instead of learning a real pattern, which would make it perform poorly on new people. To reduce this risk, for every round of our LOSO test, we first removed any measurement that barely changed across videos (since it carries little information), and then kept only the 50 measurements that a Random Forest model found most useful, based only on the training data for that round.

V. MICRO-EXPRESSION DETECTOR

Separately from the broader measurements above, we built a simple, rule-based tool whose only job is to spot actual micro-expression events, matching the textbook definition as closely as possible. While Section IV describes the movement of each face muscle using general statistical patterns calculated across the whole video, this section takes a stricter, more targeted approach. Rather than describing overall movement texture, we explicitly scan for and flag individual moments that match the precise timing definition of a micro-expression. For each of the 17 face muscles, the tool scans through the video frame by frame and looks for moments where:

- The movement score goes above a small threshold (0.15 out of 5), meaning some real movement is happening.
- This movement lasts somewhere between 1 and 6 frames, short enough to count as a micro-expression, and not a longer, more deliberate expression.
- The movement either starts suddenly (a sharp jump from the previous frame) or reaches a fairly strong peak, which helps rule out very faint, ambiguous flickers.

Every time the tool finds a movement matching all three rules, it records which face muscle was involved, how long it lasted, how strong it was at its peak, and how sudden the onset was. These events were counted up for each face muscle, separately for every video.

For each of the 17 face muscles, we computed six summary measurements (event count, mean peak intensity, maximum peak intensity, mean duration, mean onset sharpness, and event rate), yielding 103 micro-expression event measurements in total. We then used a statistical test (the Mann-Whitney U test [21]) to check, for each of these 103 measurements, whether lie videos differed from truth videos. This test does not assume the numbers are spread out in any particular shape, which makes it suitable for count data like ours. We treated any result with a p-value under 0.10 as worth a closer look, since we expected any real effect here to be small given how subtle micro-expressions are.

VI. MACHINE LEARNING MODELS

We tested three different types of machine learning models: Support Vector Machines (SVM), Random Forests, and XGBoost. SVM is a method that works well even with a small number of examples [22], [23], which suits our 320-video dataset. Random Forest builds many small decision trees and combines their votes [24], which also gives us a useful way to see which measurements mattered most. XGBoost is a more advanced method that builds trees one after another, each one correcting the mistakes from the last, and includes built-in protection against memorizing the training data too closely [25].

Before feeding our numbers into any model, we rescaled them using a method called RobustScaler, which is less affected by extreme values than standard scaling [26]. This mattered because our spike-detection numbers naturally include some large, meaningful values that we did not want the scaling step to treat as mistakes. For the SVM, we used a penalty setting of C=10. For Random Forest, we used 500 trees with a maximum depth of 10 levels. For XGBoost, we used 300 rounds of boosting with a maximum tree depth of 4. Hyperparameter values were selected through limited preliminary testing on a small number of training folds only, never using any held-out test-fold data; no nested cross-validation or exhaustive grid search was performed due to the computational cost of repeating the full 80-fold procedure for each candidate setting. This is acknowledged as a limitation in Section IX.

We report three numbers for every test: accuracy (the percentage of videos correctly labeled as lie or truth), F1 score (a balance of precision and recall, useful when checking for bias toward one label), and AUC-ROC (a number between 0 and 1 showing how well the model can rank a lie video as more "lie-like" than a truth video, regardless of any specific cutoff point). A score of 0.50 on AUC-ROC means the model is no better than a coin flip.

VII. RESULTS

Table I shows the results for every combination of machine learning model and feature set tested in this study. Each number comes from the full 80-fold Leave-One-Subject-Out testing process described in Section VI, meaning every result reflects performance on people the model never saw during training.

Using all 245 measurements without narrowing them down first, every model performed close to chance level: SVM reached 47.8%, Random Forest reached 51.3%, and XGBoost reached 52.2%. This is most likely explained by overfitting. With 245 measurements but only 320 videos, there is simply too much information relative to the number of examples, making it easy for a model to latch onto patterns that exist only by coincidence in the training data and do not hold up on new people.

Narrowing the measurements down to the 50 most useful ones, based only on the training data in each fold, improved results clearly for two of the three models. SVM improved to 55.9% and Random Forest improved to 56.3%, both slightly above chance. XGBoost, however, barely changed at all, reaching only 51.3%. A likely explanation is that XGBoost already includes its own built-in protection against overfitting (through limiting tree depth and randomly leaving out some measurements during training), so narrowing the measurements down beforehand did not add much further benefit, and may have even removed some measurements that XGBoost's own internal protections would otherwise have used to advantage.

We also tested all three models using only the 12 micro-expression event measurements described in Section V, with no other measurements included. Here, performance dropped to chance level or below for every model: SVM reached 46.9%, Random Forest reached 48.8%, and XGBoost reached 48.4%. This consistent pattern, across three different types of models strongly suggests the problem lies in the measurements themselves rather than in any one particular model's limitations.

TABLE I
RESULTS FOR EACH METHOD

Model and Feature Set	**Results**		
	Accuracy	**F1 Score**	**AUC-ROC**
SVM, all 245 features	47.8%	0.492	0.530
RF, all 245 features	51.3%	0.516	0.500
XGBoost, all 245 features	52.2%	0.517	0.512
SVM, top 50 features	55.9%	0.555	0.459
RF, top 50 features	56.3%	0.551	0.607
XGBoost, top 50	51.3%	0.519	0.517
SVM, ME events only (12)	46.9%	0.485	0.426
RF, ME events only (12)	48.8%	0.468	0.482
XGBoost, ME events only (12)	48.4%	0.480	0.474

Beyond the systematic comparison in Table I, we tested whether combining our two types of measurements (the broad statistical measurements and the targeted micro-expression event measurements) could improve on either type alone. Table II shows the results of this additional comparison, alongside the best single-type result from Table I for reference.

Our strongest result came from a Random Forest model using 53 measurements: the same top 50 statistical measurements used above, with 3 extra micro-expression event measurements added on top (the mean peak strength of AU05 events, the mean peak strength of AU02 events, and the mean duration of AU26 events). This combined model reached 58.8% accuracy, compared to 56.3% for the top-50 statistical-only model. Both results were significantly better than the 50% chance level (hybrid: $p=0.0010$; top-50 only: $p=0.0145$, one-sided binomial test). To estimate the precision of these accuracy figures, we computed fold-level accuracy across the 80 LOSO folds: the hybrid model's accuracy was 0.594 with a 95% confidence interval of [0.544, 0.645], and the top-50 model's accuracy was 0.567 with a 95% confidence interval of [0.520, 0.614]. These intervals overlap substantially. A McNemar's test directly comparing the two models' predictions on the same videos confirmed that the difference between them was not statistically significant ($p=0.332$). We therefore cannot conclude with confidence that adding the micro-expression event measurements meaningfully improved performance beyond what could be expected from random variation, although the hybrid model's higher AUC-ROC (0.642 vs. 0.607) suggests it may rank videos somewhat more reliably even where the binary classification accuracy is statistically indistinguishable.

We also tested a much smaller, hand-picked combination: just 12 manually selected statistical measurements, chosen to represent a range of different face muscles and avoid measuring the same thing twice, together with the same 3 micro-expression event measurements, for a total of 15 measurements. This smaller model reached 55.0% accuracy, only 2.2 percentage points below the 53-measurement model despite using less than a third as many measurements. This suggests that a small, easier-to-interpret set of measurements can capture most of the useful information without needing the full 53-measurement set.

TABLE III
RESULTS WHEN COMBINING STATISTICAL AND MICRO-EXPRESSION EVENT MEASUREMENTS

Model and Feature Set	**Number of Features**	**Results**		
		Accuracy	**F1 Score**	**AUC-ROC**
RF, top 50 features (statistical only)	50	56.3%	0.551	0.607
RF, combined (statistical + ME events)	53	58.8%	0.593	0.642
RF, small combined set	15	55.0%	0.544	0.593

The consistently weak performance of the micro-expression-only models in Table I raises a natural question: if these 12 measurements were so unhelpful for prediction, why did we include them in the combined model in Table II at all, and why does Section V argue that micro-expression detection found something real?

To answer this, we first looked at the overall pattern of detected micro-expression events, independent of any classifier. Averaged across all videos, lie videos contained 29.6 micro-expression events and truth videos contained 29.0 events. This near-identical overall rate indicates that deceptive speakers do not simply produce a greater number of transient facial movements in general. If a real difference exists, it must involve which specific face muscles are active, not how often the face moves overall.

When we examined individual face muscles rather than the total count, one stood out clearly. AU05 (upper eyelid raise) showed a consistent, statistically significant difference between lie and truth videos across several different measurements: more events on average (0.92 versus 0.76 per video, $p=0.039$), longer events ($p=0.013$), a sharper onset ($p=0.026$), and stronger peak intensity ($p=0.026$), all higher in lie videos. AU02 (outer brow raise) showed a similar but slightly weaker pattern ($p=0.031$ to 0.061). These results indicate genuine, statistically real differences between how lie videos and truth videos behave, on average, across the 320 videos in our dataset.

Despite this, a classifier trained using only these 12 measurements performed no better than guessing. To understand this apparent contradiction, we examined how closely related the 12 measurements were to each other, using correlation analysis. We found a severe overlap. Several AU05 measurements were correlated with each other at 0.61 to 0.99, and several AU02 measurements were correlated at 0.63 to 0.95. In other words, although we had 12 separate numbers, most of them were essentially repeating the same underlying information in slightly different forms. After accounting for this overlap, the 12 measurements carried only about 3 genuinely independent pieces of information, one related to AU05's behavior, one related to AU02's behavior, and one related to AU26's event duration.

This explains the failure clearly. A real, population-level pattern (more AU05 activity in liars, on average) does not automatically mean that pattern is strong enough, or carries enough independent information, to reliably separate individual lie videos from individual truth videos, especially once the natural differences between people's resting facial expressiveness are taken into account. Three independent pieces of information are simply not enough for any of the three models we tested to learn a reliable, generalizable rule.

This is an important and informative result rather than a disappointing one. Our rule-based detector correctly identified a genuine pattern associated with deception. However, that pattern, on its own, was too narrow and too repetitive across its own measurements to support a working lie detector by itself. It became useful only once it was combined with the much broader set of statistical measurements, as shown in Table II.

To understand what our best-performing model (the 53-feature combined Random Forest from Table II) was actually using to make its predictions, we applied a method called SHAP [27], which estimates how much each individual measurement pushed a given prediction toward "lie" or toward "truth."

The single most influential measurement was the peakiness (kurtosis) of AU26 (jaw drop), followed by how often AU17 (chin raise) was active, and how variable AU14 (mouth dimple area) was throughout the video. In each of these cases, higher values pushed the model's prediction toward "lie." Measurements describing the shape of a movement pattern over time (specifically peakiness and lopsidedness, rather than simple averages or maximum values) appeared far more often among the most influential measurements overall. This supports the central idea behind our feature design in Section IV: a face muscle that moves in occasional sharp spikes, rather than smoothly and steadily, carries more useful information for telling lies apart from truths.

Notably, AU05 also appeared multiple times among the most influential measurements identified by SHAP, despite being one of the measurements that failed when tested entirely on its own in Section VII-C. This is a meaningful point of agreement between our two separate analysis methods: both the rule-based event detector and the SHAP analysis of the machine learning model independently pointed to AU05 as relevant to lying, even though neither approach alone was sufficient to reliably predict it.

VIII. DISCUSSION

Our best result of 58.8% accuracy is not impressive in absolute terms, but it needs to be judged against the right comparison point. Humans correctly detect lies only about 54% of the time on average [1], so a computer reaching 58.8% under a fair, person-independent test, a result significantly above chance ($p=0.001$), represents a small but genuine improvement over unaided human judgment.

It is worth comparing this to other published results on related deception detection tasks. Wu et al. [16] and Pérez-Rosas et al. [17] both report considerably higher accuracy (up to 0.877 AUC and 60-75% accuracy respectively), but in both cases their models used more than facial movement alone. Wu et al. combined facial features with audio and text, and Pérez-Rosas et al. combined facial movement with other verbal and non-verbal cues. Neither study isolates facial movement on its own the way the present paper does, so their higher numbers cannot be directly attributed to facial cues alone, and a fully controlled, side-by-side comparison was outside the scope of this paper. Still, the pattern is suggestive: studies that combine multiple types of cues tend to report noticeably higher accuracy than studies, including ours, that rely on facial movement by itself. This is consistent with the conclusion we draw below: facial movement alone, even when measured carefully, may simply not carry enough information on its own to reliably detect lying, and combining it with other channels such as voice or word choice appears necessary for stronger performance.

The main question this paper set out to answer was whether micro-expressions, used on their own and without any other supporting facial or behavioral measurements, are sufficient to reliably detect lying. Our results give a clear answer: they are not, at least not with current detection methods and dataset sizes.

This finding warrants explicit statement, because it stands in some tension with the influential idea, first proposed by Ekman [2], that brief involuntary facial movements reliably "leak" concealed emotion during deception. Our findings do not contradict the idea that something real is happening. We found a genuine, statistically significant pattern in which AU05 (upper eyelid raise) occurred more often during lying than during truth-telling, and this pattern was detected independently by two separate methods, our rule-based event detector and the SHAP analysis of our machine learning model. This convergence gives us reasonable confidence that the AU05 pattern itself is real rather than a coincidence of our particular dataset.

However, finding a real pattern across many people is not the same as having a pattern strong enough to predict any one individual case reliably. The size of the average difference we found was small (less than one extra AU05 event per video, on average, between lie and truth videos), and people naturally differ enormously from each other in how expressive their faces are in general. This natural between-person variation appears to be large enough to mostly drown out the deception-related signal when trying to classify a single new, previously unseen person. Our correlation analysis added a further, more technical explanation: the handful of micro-expression measurements that did show a significant pattern were themselves highly repetitive, providing only about three genuinely independent pieces of information in total, which is too little for any classifier to learn a reliable, generalizable rule from.

Taken together, this paints a more nuanced picture than either a simple confirmation or a simple rejection of Ekman's theory. Micro-expressions may well leak real information about concealed emotion, consistent with the original theory, but our results suggest this leakage is too weak and too narrow, at least as captured by current automatic detection methods, to serve as a standalone, reliable tool for catching individual lies. This suggests the practical value of micro-expression theory for automated lie detection may depend on combining it with other signals, rather than relying on it in isolation.

While the combined model numerically outperformed the statistical-only model, this difference was not statistically significant (Section VII), so this result should be interpreted cautiously rather than as firm evidence that combining the two measurement types reliably improves performance. What can be said with more confidence is that the micro-expression event measurements, on their own, were clearly insufficient (Section VII), while the broader statistical measurements carried most of the usable signal in this study. This suggests that the most promising path forward is not to treat micro-expression detection as a complete solution, but as one useful ingredient among several.

IX. LIMITATIONS AND FUTURE WORK

Several limitations should be kept in mind when interpreting these results. The lies in our dataset were told in a low-pressure laboratory setting, where participants made true or false statements about their opinions of other people, rather than in a high-pressure, real-world setting such as a criminal interrogation. The emotional and physical stakes of lying in our dataset are likely much lower than in many real-world situations, and it is not yet known whether our findings, including the AU05 pattern, would hold up under higher-stakes conditions where fear of being caught might be considerably stronger.

OpenFace, the tool used to measure facial movement in this study, is not perfectly accurate and can make small measurement errors, particularly when a person's head is turned away from the camera or lighting conditions are poor. This kind of measurement noise could be partly hiding a true pattern that is somewhat stronger than what we were able to detect. Our AU05 finding ($p=0.039$) was identified through an exploratory screening across all 103 micro-expression event measurements, which increases the chance that it could be a false positive appearing by chance alone. For this reason, it should be treated as a promising lead for future research rather than a fully confirmed result until it has been replicated using new, independent data.

Hyperparameters were not tuned using nested cross-validation, meaning the reported performance may be marginally optimistic compared to a fully nested approach, although no test-fold information was used during selection.

Random Forest classification involves some inherent randomness even with a fixed random seed in certain configurations, and small variations in reported accuracy (on the order of 1-2 percentage points) between repeated runs of the same model and feature set should be expected and do not indicate a meaningful difference in performance.

The sample size of 80 people, while reasonable for this type of research, is still relatively small for a person-independent testing approach, where each test fold only includes four videos from a single new person. A larger dataset would allow for a more precise and stable estimate of how well these methods actually perform.

Future work could take several directions. Testing deep learning models that learn directly from raw video, without manually designed measurements, is one possibility; although such models typically require far more training data than the 320 videos available here. Combining facial measurements with voice tone and word-choice patterns is a more immediately promising direction, since lying is a behavior that likely shows up across multiple communication channels at once, not just the face. Our discussion in Section VIII suggests this combined approach may be necessary to substantially improve on the results reported here. Finally, testing these same methods on higher-stakes, real-world lying situations, rather than a controlled laboratory setting, would help confirm whether the patterns identified in this paper, particularly the AU05 finding, generalize beyond the specific conditions of our dataset.

X. CONCLUSIONS

This paper asked a focused question: can facial micro-expressions, used entirely on their own, reliably tell whether someone is lying? Using a public dataset of 320 videos and a strict person-independent testing method, in which the model was always tested on people it had never seen during training, we found that the answer is no, not with current detection methods. Our best model, which combined micro-expression measurements with broader statistical measurements of facial movement, reached 58.8% accuracy, a modest but statistically significant improvement over chance ($p=0.001$) and over the human baseline of 54%, although this model was not found to significantly outperform a comparable model using only the broader statistical measurements. However, a model built using only the strict, theory-driven micro-expression measurements performed no better than random guessing, despite those same measurements showing a real, statistically significant link between one specific eye movement (AU05) and lying.

This combination of results tells a coherent and important story. Ekman's theory that involuntary facial leakage occurs during deception appears to hold up to some degree, since we did find a real, consistent pattern connected to deception. However, this pattern alone is too weak, too narrow, and too repetitive across its own measurements to support reliable lie detection by itself. When compared with other published work that combines facial movement with voice and word-choice information and tends to report higher accuracy, our results support the broader conclusion that facial micro-expressions are likely a genuine but incomplete piece of the deception puzzle. They appear better suited to confirming theoretical patterns about deception across groups of people than to serving as a standalone practical tool for catching lies in any one individual. Future systems aiming for reliable, real-world lie detection will likely need to combine micro-expression analysis with other behavioral and physiological signals, rather than relying on facial movement alone.

ACKNOWLEDGEMENT

The authors thank the creators of the MU3D dataset for making this research possible.

REFERENCES

[1]. Bond Jr, C. F., & DePaulo, B. M. (2006). Accuracy of deception judgments. Personality and social psychology Review, 10(3), 214-234.

[2]. Ekman, P. (2009). Lie catching and microexpressions. The philosophy of deception, 1(2), 5.

[3]. Frank, M. G., & Svetieva, E. (2014). Microexpressions and deception. In Understanding facial expressions in communication: Cross-cultural and multidisciplinary perspectives (pp. 227-242). New Delhi: Springer India.

[4]. Zhao, G., Li, X., Li, Y., & Pietikäinen, M. (2023). Facial micro-expressions: An overview. Proceedings of the IEEE, 111(10), 1215-1235.

[5]. Yan, W. J., Wu, Q., Liang, J., Chen, Y. H., & Fu, X. (2013). How fast are the leaked facial expressions: The duration of micro-expressions. Journal of nonverbal behavior, 37(4), 217-230.

[6]. Baltrušaitis, T., Robinson, P., & Morency, L. P. (2016, March). Openface: an open source facial behavior analysis toolkit. In 2016 IEEE winter conference on applications of computer vision (WACV) (pp. 1-10). IEEE.
[7]. Amos, B., Ludwiczuk, B., & Satyanarayanan, M. (2016). Openface: A general-purpose face recognition library with mobile applications. CMU School of Computer Science, 6(2), 20.
[8]. Ekman, P., & Friesen, W. V. (1978). Facial action coding system. Environmental Psychology & Nonverbal Behavior.
[9]. Ning, M., Salah, A. A., & Ertugrul, I. O. (2024). Representation learning and identity adversarial training for facial behavior understanding. arXiv preprint arXiv:2407.11243.
[10]. Frank, M. G., & Ekman, P. (1997). The ability to detect deceit generalizes across different types of high-stake lies. Journal of personality and social psychology, 72(6), 1429.
[11]. Jordan, S., Brimbal, L., Wallace, D. B., Kassin, S. M., Hartwig, M., & Street, C. N. (2019). A test of the micro- expressions training tool: Does it improve lie detection?. Journal of Investigative Psychology and Offender Profiling, 16(3), 222-235.
[12]. Coan, J. A., & Gottman, J. M. (2007). The specific affect coding system (SPAFF). Handbook of emotion elicitation and assessment, 267, 285.
[13]. Hopf, A. G., Buchheim, A., Hopf, M., & Eilert, D. W. (2026). A Psychological Guide to Lower Face Botulinum Toxin Injections: Baseline Emotional Functions of Facial Expressions. Journal of Cosmetic Dermatology, 25(4), e70833.
[14]. Namba, S., Kabir, R. S., Miyatani, M., & Nakao, T. (2017). Spontaneous facial actions map onto emotional experiences in a non-social context: toward a component-based approach. Frontiers in Psychology, 8, 633.
[15]. Bress, K. S., & Cascio, C. J. (2024). Sensorimotor regulation of facial expression–an untouched frontier. Neuroscience & Biobehavioral Reviews, 162, 105684.
[16]. Wu, Z., Singh, B., Davis, L., & Subrahmanian, V. (2018, April). Deception detection in videos. In Proceedings of the AAAI conference on artificial intelligence (Vol. 32, No. 1).
[17]. Pérez-Rosas, V., Abouelenien, M., Mihalcea, R., & Burzo, M. (2015, November). Deception detection using real-life trial data. In Proceedings of the 2015 ACM on international conference on multimodal interaction (pp. 59-66).
[18]. Dinges, L., Fiedler, M. A., Al-Hamadi, A., Hempel, T., Abdelrahman, A., Weimann, J., & Bershadskyy, D. (2023). Automated deception detection from videos: using end-to-end learning based high-level features and classification approaches. arXiv preprint arXiv:2307.06625.
[19]. Lloyd, E. P., Deska, J. C., Hugenberg, K., McConnell, A. R., Humphrey, B., & Kunstman, J. W. (2017). Miami University deception detection video database. Manuscript under review.
[20]. Esterman, M., Tamber-Rosenau, B. J., Chiu, Y. C., & Yantis, S. (2010). Avoiding non-independence in fMRI data analysis: leave one subject out. Neuroimage, 50(2), 572-576.
[21]. MacFarland, T. W., & Yates, J. M. (2016). Mann–whitney u test. In Introduction to nonparametric statistics for the biological sciences using R (pp. 103-132). Cham: Springer International Publishing.
[22]. Hearst, M. A., Dumais, S. T., Osuna, E., Platt, J., & Scholkopf, B. (1998). Support vector machines. IEEE Intelligent Systems and their applications, 13(4), 18-28.
[23]. Mammone, A., Turchi, M., & Cristianini, N. (2009). Support vector machines. Wiley Interdisciplinary Reviews: Computational Statistics, 1(3), 283-289.
[24]. Salman, H. A., Kalakech, A., & Steiti, A. (2024). Random forest algorithm overview. Babylonian Journal of Machine Learning, 2024, 69-79.
[25]. Bentéjac, C., Csörgő, A., & Martínez-Muñoz, G. (2019). A comparative analysis of XGBoost. arXiv preprint arXiv:1911.01914.
[26]. Qian, H., Wen, Q., Sun, L., Gu, J., Niu, Q., & Tang, Z. (2022, May). Robustscaler: Qos-aware autoscaling for complex workloads. In 2022 IEEE 38th International Conference on Data Engineering (ICDE) (pp. 2762-2775). IEEE.
[27]. Ponce- Bobadilla, A. V., Schmitt, V., Maier, C. S., Mensing, S., & Stodtmann, S. (2024). Practical guide to SHAP analysis: Explaining supervised machine learning model predictions in drug development. Clinical and translational science, 17(11), e70056.

